Silvano Salvador

Toponomastic Networks

Information Geometry, Stochastic Processes, and Spectral Structure of Phonotactic Substrate Relations

Volume II

In Three Parts:

Part I — Information Geometry and Topological Data Analysis
Part II — Stochastic Processes and Markov Chains
Part III — Spectral and Arithmetic Structure

A Companion Series to Toponomastic Entropy

Contents

Preface **vii**

I Information Geometry and Topological Data Analysis **1**

Preface to Part I **2**

1 From a Graph of Zones to a Manifold of Toponyms **4**
1.1 What Volume I Could Not See 4
1.2 Why Linear PCA Is the Wrong Default for Simplex Data 5

2 The Fisher–Rao Information Metric **6**
2.1 The Fisher Information Metric on a Parametric Family 6
2.2 The Multinomial Simplex as a Round Sphere 7
2.3 Relation to Other Information-Theoretic Quantities 7

3 The Statistical Manifold of a Substrate Zone **9**
3.1 A Declared Synthetic Construction 9
3.2 Fisher–Rao versus Euclidean Distance on the Same Cloud 10

4 Geodesics and Curvature on the Statistical Manifold **12**
4.1 From a Distance Formula to a Geometry 12
4.2 The Fisher Information Metric in Coordinates 12
4.3 Curvature: Why the Simplex Is Not Flat 13
4.4 The Fréchet Mean: Algorithm and an Explicit Computation 14
4.5 The Geodesic Equation and Explicit Interpolation via Spherical Linear Interpolation 16

5 Literature Review: Information Geometry in Linguistics, Phylogenetics, and Genomics **19**
5.1 Why a Literature Review Belongs Here, Not in the Preface 19
5.2 The Fisher–Rao Metric Outside Information Geometry Proper 19
5.3 Persistent Homology Outside Pure Topology 21
5.4 What This Comparison Does and Does Not License 22

6 Topological Data Analysis: Persistent Homology **23**
6.1 From a Point Cloud to a Filtration of Simplicial Complexes 23
6.2 A Worked Toy Example 24

7 Applying the Pipeline to the Three Representative Zones **26**
7.1 Persistence Results 26
7.2 Sensitivity to the Synthetic Construction 27

8 Real-Data Validation: Persistent Homology and Geometry on Actual GeoNames Toponym Clouds **28**
8.1 Zone Definitions on Real GeoNames Data 28
8.2 Aggregate Entropy: Real Data Against the Published Benchmark . . 29
8.3 Closing the Basque Gap: Adding the French Departments 30
8.4 Toponym-Level Point Clouds: Real Bigram Vectors 31
8.5 The Fréchet Mean on Real Data, and a Predicted Artefact Confirmed 32
8.6 Persistent Homology on Real Toponym Clouds 33
8.7 Fisher–Rao Versus Euclidean PCA, Real Data 34
8.8 Declared Simplifications and Threats to Validity 35

9 Extending Real-Data Validation: Seventeen New National Corpora, Eighteen Country-Level Pools **37**
9.1 Pool Construction 37
9.2 Toponym-Level Point Clouds and Fisher–Rao Distances 39
9.3 The Fréchet Mean on the Eighteen Additional Clouds 40
9.4 Discussion: What Twenty-One Real National Pools Show, and Do Not Show 41
9.5 Declared Simplifications and Threats to Validity, Extended 43

10 Stability Theorems for Persistent Homology **44**
10.1 Why a Pipeline Needs a Stability Guarantee 44
10.2 Persistence Diagrams as a Metric Space 44
10.3 The Stability Theorem 45
10.4 A Coordinate-Free Strengthening: Stability Under Gromov–Hausdorff Perturbation 46
10.5 What Stability Licenses for This Volume's Pipeline 47
10.6 An Empirical Resampling Experiment: Estimating the Noise Floor . . 48

11 Synthesis and Critical Comparison: Fisher–Rao Geometry versus Euclidean PCA **50**
11.1 What This Part Adds to the Series 50
11.2 A Closer Look at the Distortion Claim 50
11.3 The Central Limitation, Restated 52
11.4 Agenda for Part II 53

12 Reproducibility: The Synthetic Data Generation and Persistence Code **54**

13 Reproducibility: The Extended Real-Data Pipeline **57**

14 Real Persistent Homology Across Twenty-One National and Zone Pools **61**
14.1 The Twenty-One Pools 62
14.2 Persistence Diagrams: The Most and Least Structured Pools 63

15 Inter-Pool Clustering: Real Hierarchical Structure from Aggregate Bigram Divergence **65**
15.1 The Pairwise Divergence Matrix 65
15.2 Hierarchical Clustering and Its Cophenetic Fidelity 66

16 Within-Country Structure: Admin1-Level Entropy Variance Against a Real Permutation Null **69**
16.1 Method: A Real Permutation Test 69
16.2 Results 70

17 Extending the Admin1 Permutation Test: Sixteen Further National Pools **73**
17.1 Method 73
17.2 Results: Six Countries Pass 74
17.3 The Ten Countries That Could Not Be Tested 75
17.4 Updated Synthesis 76

18 Resolving the Deferred Items: A Closing Assessment **77**

19 Reproducibility: The Real Persistent-Homology, Clustering, and Admin1-Variance Pipeline **79**

II Stochastic Processes and Markov Chains **82**

Preface to Part II **83**

20 From a Synchronic Graph to a Diachronic Process **85**
20.1 What a Static Backbone Cannot, by Itself, Say About History 85
20.2 Why the Backbone, Not the Full Graph 85
20.3 Two Formalisms, One Backbone: A Preview 86

21 Random Walks on Weighted Graphs **87**
21.1 Transition Matrix and Reversibility 87
21.2 First-Passage (Hitting) Times 88
21.3 Spectral Gap and Mixing 89

22 Markov Chain Convergence: Mixing Times and Spectral Gaps **90**
22.1 Why Convergence Needs Its Own Chapter 90
22.2 Total Variation Distance and the Convergence Theorem 90
22.3 The Variational Characterisation of the Spectral Gap 92
22.4 Why a Spanning Tree Mixes Slowly: A Lower Bound via Cheeger's Inequality 93
22.5 What Slow Mixing Does and Does Not Say About the Series' Findings 95

23 Literature Review: Diffusion Models in Dialectometry and Historical Linguistics **96**
23.1 Why a Literature Review Belongs Here 96
23.2 The Wave Model and Its Successors 96
23.3 The Gravity Model of Linguistic Diffusion 97

23.4 Diffusion and Drift in Dialectometry . . . 98
23.5 Diffusion Models from Historical and Population Linguistics . . . 98
23.6 Spatial Diffusion of Innovations: The Hägerstrand Tradition . . . 99
23.7 Random Walks, Effective Resistance, and the General Probabilistic Literature . . . 100
23.8 A Synthesising Reference Point: Rogers' Diffusion of Innovations . . . 100
23.9 What the Comparison Licenses and Does Not . . . 101

24 Reconstructing the Diffusion Backbone **102**
24.1 What Is Real and What Is Reconstructed . . . 102
24.2 Why a Uniform Draw Within the Real Range, and Not Some Other Distribution . . . 104
24.3 Sensitivity Check: Repeating the Reconstruction . . . 105

25 The Random Walk on the Reconstructed Backbone: Results **106**
25.1 Stationary Distribution . . . 106
25.2 First-Passage Times . . . 108
25.3 Spectral Gap . . . 109

26 What a Step Count Can and Cannot Say About History **110**
26.1 The Honest Reading of Table 25.2 . . . 110
26.2 Point Estimates Versus Robust Rankings . . . 111

27 Electrical Network Theory and Effective Resistance **112**
27.1 A Second Lens on the Same Backbone . . . 112
27.2 The Resistor-Network Correspondence . . . 112
27.3 Effective Resistance for the Five Representative Pairs . . . 114
27.4 Effective Resistance and the Tree Structure . . . 115
27.5 Foster's Theorem: A Built-In Consistency Check . . . 115
27.6 Effective Resistance as a Metric . . . 116
27.7 Limits of the Cross-Check . . . 117
27.8 What This Volume Adds . . . 118
27.9 The Central Limitation, Restated . . . 118
27.10Two Methodological Lessons for the Series as a Whole . . . 119
27.11Agenda for Part III . . . 119

28 Reproducibility: Backbone Reconstruction and Random-Walk Code **121**

III Spectral and Arithmetic Structure 122

Preface to Part III **123**

29 From Graphs and Walks to Spectra and Sequences **125**
29.1 What Entropy Cannot See . . . 125
29.2 What Volume I's Laplacian Spectrum Has Not Yet Been Asked . . . 126

30 The DFT in Full: Parseval, Leakage, and the Fisher g-Test **127**
30.1 Definitions . . . 127
30.2 Parseval's Theorem: The Periodogram as a Variance Decomposition . 128
30.3 Spectral Leakage: What Happens When the Period Does Not Divide N 129
30.4 Significance Testing: Fisher's g-Statistic . . . 130
30.5 The Exact Null Distribution of $g_{\max}$ . . . 131
30.6 An Alternative Correction: Controlling the False Discovery Rate . . . 132
30.7 The DFT as a Graph Fourier Transform on a Cycle . . . 133

31 Literature Review: Spectral Methods in Sequence and Phonological Analysis **135**
31.1 Spectral Periodicity Detection in Biological Sequence Data . . . 135
31.2 Classical Time-Series Spectral Analysis . . . 136
31.3 Rhythm and Periodicity in the Speech Signal: A Different Question . 137
31.4 Random Matrix Theory in Network Science . . . 137
31.5 Multitaper and Wavelet Alternatives to the Raw Periodogram . . . 138
31.6 Symbolic, Non-Acoustic Periodicity in Written Language . . . 138

32 Constructing Phonotactic Rank Sequences **140**
32.1 Construction Method . . . 140
32.2 A Worked Example . . . 141
32.3 Summary of the Injection Design . . . 142

33 Spectral Results: Recovering the Injected Periodicities **143**
33.1 Periodogram and Fisher Test, All Thirty-Three Zones . . . 143
33.2 The Two Missed Injections . . . 144
33.3 The One False Positive . . . 144
33.4 Period Recovery Accuracy . . . 145

34 Random Matrix Theory and the Null Spectrum **146**
34.1 The Gaussian Orthogonal Ensemble . . . 146
34.2 The Wigner Semicircle Law . . . 147
34.3 Deriving the Joint Eigenvalue Density: Where the Vandermonde Factor Comes From . . . 148
34.4 Level Repulsion and the Spacing-Ratio Statistic . . . 150
34.5 The Laplacian Constraint: Why $\mathcal{G}$ Is Not, Even in Principle, a GOE Matrix . . . 151

35 Volume I's Laplacian Spectrum, Revisited **153**
35.1 The Full Spectrum as a Distribution . . . 153
35.2 Spacing Statistics . . . 154
35.3 Permutation Test: Is μ_2 Unusual? . . . 154

36 What a Spectral Peak Can and Cannot Say **156**
36.1 The Two Findings Are Not Symmetric . . . 156
36.2 If Real Sequence Data Became Available . . . 156
36.3 Why the Null Laplacian Result Is the More Valuable of the Two . . . 157
36.4 A Further Caveat: Significance Is Not the Same Question as Size . . . 157

37 Synthesis: Closing the Toponomastic Networks Tetralogy **159**
37.1 What Four Volumes Have Established, in One Place 159
37.2 One Thread Running Through All Four Volumes 160
37.3 What This Series Has Not Done 160
37.4 A Methodological Lesson for Future Multi-Method Studies 160
37.5 An Open Door, Not a Promise 161

38 Reproducibility: Sequence Generation, Spectral Tests, and Permutation Null **162**

Bibliography **164**

Preface

This volume continues *Toponomastic Networks* from its completed first volume — the graph-theoretic foundations built on the thirty-three-zone substrate backbone — and develops three further mathematical perspectives on the same underlying system, gathered here as a single coherent book in three parts rather than as separate volumes.

Part I descends from the zone-level graph to the level of individual toponyms, treating each zone as a cloud of points on the Fisher–Rao statistical manifold and applying persistent homology to ask what shape, if any, these clouds have. Part II introduces the dimension absent from both the original graph and Part I's manifold: time, via random walks and Markov chain theory on the inherited backbone. Part III asks whether the fine-grained *order* of phonotactic units, not merely their aggregate frequencies, carries detectable periodic structure, using the discrete Fourier transform and revisits the original graph's Laplacian spectrum against a declared null model.

Each part carries forward its own, increasingly explicit data-honesty caveat, stated in full at the head of its own preface below: what is real, inherited, reconstructed, or declared synthetic is flagged at every point where it materially affects a numerical result. No claim in this volume should be read as an empirical finding about real historical relationships beyond what *Toponomastic Entropy*'s original entropy and KL-divergence computations already established; the mathematical apparatus developed in each part is fully general and applies unchanged to real data whenever such data becomes available within this series' scope.

Part I

Information Geometry and Topological Data Analysis

Preface to Part I

Volume I of this series treated each of the thirty-three substrate zones as a single point — a fitted bigram distribution — and asked what relational structure emerges when all thirty-three such points are considered simultaneously as the vertex set of a graph. The present volume descends one level of resolution. Instead of one point per zone, we now consider the cloud of *individual toponyms* that, aggregated and smoothed, produced each zone's distribution in the first place. Every toponym, once converted to a bigram frequency vector, is itself a point on the probability simplex Δ^{K-1}; a zone is not one point but a *cloud* of several thousand such points, and the entropy values and KL divergences of *Toponomastic Entropy* are summary statistics – means and spreads – of clouds whose internal shape has never, in either series so far, been examined directly.

Two mathematical decisions follow from this change of resolution. First, the simplex Δ^{K-1} is not a Euclidean space: it is a curved statistical manifold, and the natural metric on a space of probability distributions is not the Euclidean distance used implicitly by ordinary PCA but the *Fisher–Rao information metric*, the unique (up to scale) Riemannian metric invariant under sufficient statistics (Čencov, 1982; Amari and Nagaoka, 2000). Treating toponym clouds as living on this curved manifold, rather than flattening them into $\mathbb{R}^K$ by fiat, is the organising idea of Chapters 2 and 3. Second, once a faithful (geodesic) notion of distance between toponyms is available, the cloud's *shape* — connected components, loops, voids — becomes a legitimate object of study in its own right, via persistent homology, the subject of Chapters 6 and 7.

A second, sharper methodological caveat than Volume I's. Volume I's proxy distance was a declared simplification of a quantity that, in principle, *could* be measured directly (a cross-corpus KL divergence) from data already collected by the parent series. The present volume faces a more fundamental data gap: *Toponomastic Entropy* publishes per-zone aggregate statistics (entropy, KL divergence, sample size) but not the raw toponym-by-toponym bigram count vectors from which those aggregates were computed, and those raw vectors are not available within the scope of this volume either. Every toponym-level point cloud analysed in Chapters 3 and 7 is therefore a *declared synthetic construction*: a calibrated Dirichlet–multinomial generative model, fitted so that its expected bias-corrected Shannon entropy matches the real, published H_2 value of the zone it represents (Vol. I, Table 4.1), and then sampled to produce a synthetic cloud of toponym-level bigram vectors. This is the same honesty principle Volume I applied to its inter-regional proxy edges, pushed one level further: every synthetic number in this volume is reproducible from the generative script documented in Appendix 12, is explicitly flagged wherever it appears, and is never presented as a measurement of real toponymic data. The mathematical ap-

paratus developed here — the Fisher–Rao metric, the persistent homology pipeline, the comparison with linear PCA — is fully general and applies unchanged to real toponym-level bigram vectors whenever those become available; what this volume cannot do, and does not claim to do, is report what shape the *real* toponym clouds of *Toponomastic Entropy* actually have.

Relation to the rest of the series. This volume uses, without modification, the thirty-three-zone vertex list and bias-corrected Shannon entropies of Volume I, Table 4.1 (itself inherited from *Toponomastic Entropy*, Vol. I, Table 11.1). Three zones spanning the observed entropy range — Basque ($H_2 = 7.652$, the most diverse zone in Iberia), Armenian ($H_2 = 7.540$, the most isolated zone of Anatolia–Caucasus by KL divergence), and Libya ($H_2 = 6.655$, the least diverse zone in the entire thirty-three-zone system) — are carried through every worked example as a representative low/medium/high contrast triple; the full calibration procedure generalises to all thirty-three zones and is stated in general form throughout.

Chapter 1

From a Graph of Zones to a Manifold of Toponyms

1.1 What Volume I Could Not See

Volume I's graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has $|\mathcal{V}| = 33$ vertices, one per zone, and its construction (Vol. I, Ch. 3) compresses every zone's bigram distribution $\hat{p}_i$ down to a single scalar, the bias-corrected entropy $H_2(i)$, for every inter-regional comparison, and to a single pairwise scalar $\delta(i, j)$ for every intra-regional comparison. This compression is exactly what made the graph tractable, but it is also exactly what makes a class of questions invisible to it by construction.

Remark: Three Questions a Thirty-Three-Vertex Graph Cannot Pose

1. *Is a zone's bigram distribution internally homogeneous, or does it conceal subpopulations?* Two zones with identical entropy H_2 can differ entirely in whether their toponyms form one cluster or several – entropy is blind to multimodality.
2. *Does the toponym cloud of a zone have non-trivial topology – loops or voids – that a linear summary cannot detect?* A ring-shaped cloud and a solid disc of the same diameter have identical first and second moments but different homotopy type.
3. *Is linear PCA, used as the default dimensionality-reduction tool throughout Toponomastic Entropy whenever a visualisation of zone structure is needed, actually appropriate for data that lives on a curved probability simplex rather than a flat Euclidean space?*

Answering any of these requires examining individual toponyms, not zone-level aggregates, and requires a metric appropriate to the simplex, not the ambient Euclidean coordinates PCA assumes.

1.2 Why Linear PCA Is the Wrong Default for Simplex Data

A toponym's bigram-frequency vector $q = (q_1, \ldots, q_K)$, $q_k \geq 0$, $\sum_k q_k = 1$, is a point on the probability simplex $\Delta^{K-1} = \{q \in \mathbb{R}^K : q_k \geq 0,\ \sum_k q_k = 1\}$, a $(K-1)$-dimensional curved subset of $\mathbb{R}^K$. Ordinary PCA, applied directly to the vectors q, implicitly treats Δ^{K-1} as a flat Euclidean affine subspace and measures distance as $\|q - q'\|_2$. Three well-known pathologies follow from this mismatch (Aitchison, 1986; Pennec et al., 2020):

- **Boundary distortion.** Euclidean distance between two sparse vectors near a face of the simplex (toponyms whose bigram mass concentrates on very few types) understates how different their *generative processes* are, because the simplex's natural geometry stretches near the boundary – exactly where Libya's low-entropy, concentrated toponyms (Vol. I, Table 4.1) live.
- **Scale non-invariance.** A unit Euclidean step near the centre of Δ^{K-1} (uniform-ish distributions, where Basque's high-entropy toponyms cluster) corresponds to a much smaller change in the underlying multinomial sampling distribution than the same Euclidean step near a vertex, so a single global linear projection cannot be a faithful low-dimensional summary across zones of very different entropy.
- **Linear subspaces are the wrong model class.** PCA seeks the best-fitting *affine* subspace; if the true low-dimensional structure of a toponym cloud is curved (e.g. toponyms organised along a circular alternation of two phonotactic patterns), PCA will report spuriously many components needed to explain the variance, understating how compressible the data actually is in its own natural geometry.

Methodological Transparency: This Is a Stated Limitation of the Parent Series, Not an Accusation

Toponomastic Entropy never claims its (limited) use of linear projections is metric-faithful; the parent series' primary tools are entropy and KL divergence, which are already information-theoretically appropriate. The point of this chapter is narrower: *wherever* a future volume of either series wishes to visualise or cluster toponym-level data geometrically, the Fisher–Rao metric developed in Chapter 2 is the metric that should be used, and ordinary Euclidean PCA, while a reasonable null comparison, should not be the default.

Chapter 2

The Fisher–Rao Information Metric

2.1 The Fisher Information Metric on a Parametric Family

Definition: Fisher Information Metric

Let $\{p_\theta\}_{\theta\in\Theta}$ be a smooth parametric family of probability distributions, $\Theta \subseteq \mathbb{R}^d$ open. The Fisher information matrix at θ is

$$\mathcal{I}(\theta)_{ij} \;=\; \mathbb{E}_{p_\theta}\left[\frac{\partial \log p_\theta(x)}{\partial \theta_i}\,\frac{\partial \log p_\theta(x)}{\partial \theta_j}\right].$$

$\mathcal{I}(\theta)$ is symmetric positive semi-definite and, under mild regularity, positive definite; it defines a Riemannian metric on Θ, the *Fisher–Rao metric*, via $\langle u, v\rangle_\theta = u^\top \mathcal{I}(\theta)\, v$ for tangent vectors $u, v \in \mathbb{R}^d$.

Theorem: Čencov's Characterisation (,)

The Fisher–Rao metric is, up to a constant scale factor, the *unique* Riemannian metric on a statistical manifold that is invariant under sufficient statistics (i.e. under any map that does not discard information relevant to θ). No other choice of metric on a space of probability distributions has this invariance property.

This is the precise sense in which Fisher–Rao, not Euclidean distance on raw frequency vectors, is the metric *intrinsic* to the simplex of bigram distributions: it is the metric that does not depend on an arbitrary choice of coordinates for the bigram alphabet.

2.2 The Multinomial Simplex as a Round Sphere

Proposition: The Square-Root Embedding Is a Fisher–Rao Isometry

For the multinomial family on K categories, parametrised by $q \in \Delta^{K-1}$, the map

$$\phi : \Delta^{K-1} \longrightarrow S_+^{K-1}(2), \qquad \phi(q) = 2\sqrt{q} = \left(2\sqrt{q_1}, \ldots, 2\sqrt{q_K}\right)$$

sends the simplex (with the Fisher–Rao metric) isometrically onto the positive orthant of the sphere of radius 2 in $\mathbb{R}^K$ (with the Euclidean-induced round metric). Consequently the Fisher–Rao geodesic distance between $q, q' \in \Delta^{K-1}$ equals the great-circle distance between their images:

$$d_{\mathrm{FR}}(q, q') \;=\; 2 \arccos\left(\sum_{k=1}^{K} \sqrt{q_k \, q'_k} \right).$$

Remark: Why This Matters Computationally

Proposition 2.2 converts an otherwise intractable Riemannian-geometry computation (solving the geodesic equation on a general statistical manifold) into a single inner product after a square-root transform – the entire apparatus of this volume rests on this one closed-form identity, due in essence to Bhattacharyya's coefficient $\mathrm{BC}(q, q') = \sum_k \sqrt{q_k q'_k}$, here reinterpreted geometrically rather than as a mere similarity score (Kass, 1989; Amari and Nagaoka, 2000).

Example: Bounding the Fisher–Rao Distance

Since $\mathrm{BC}(q, q') \in [0, 1]$ by the Cauchy–Schwarz inequality applied to the unit vectors $\sqrt{q}, \sqrt{q'}$, we have $d_{\mathrm{FR}}(q, q') \in [0, \pi]$, with $d_{\mathrm{FR}} = 0$ iff $q = q'$ and $d_{\mathrm{FR}} = \pi$ iff q, q' have disjoint support. This is a genuine, bounded geodesic distance, unlike the unbounded and asymmetric KL divergence used throughout *Toponomastic Entropy* and Volume I – a second methodological advantage of working on the manifold directly: d_{FR} is a true metric (symmetric, satisfies the triangle inequality), where D_{KL} is neither.

2.3 Relation to Other Information-Theoretic Quantities

Proposition: Fisher–Rao, Hellinger, and the Local KL Quadratic Form

For $q' = q + \varepsilon u$ with u a tangent direction and $\varepsilon \to 0$,

$$D_{\mathrm{KL}}(q \,\|\, q') \;=\; \tfrac{1}{2}\varepsilon^2\, u^\top \mathcal{I}(q)\, u \;+\; O(\varepsilon^3),$$

i.e. KL divergence is, to second order, exactly the squared Fisher–Rao infinitesimal distance: the Fisher–Rao metric is the *Hessian* of KL divergence at coincidence. Globally (not just infinitesimally), the Hellinger distance $\mathrm{Hel}(q, q') = \big(\sum_k (\sqrt{q_k} - \sqrt{q'_k})^2\big)^{1/2}$ and the Fisher–Rao distance are related exactly by $\mathrm{Hel}(q, q') = 2\sin\!\big(d_{\mathrm{FR}}(q, q')/4\big)$.

This connects the present volume directly back to the parent series: D_{KL}, the workhorse of *Toponomastic Entropy*, is recovered as the local quadratic approximation to the global geodesic distance introduced here. Nothing in Volumes I–VII of *Toponomastic Entropy* is contradicted by this volume; rather, the present chapter exhibits the metric of which the parent series' local, pairwise tool is the infinitesimal shadow.

Chapter 3

The Statistical Manifold of a Substrate Zone

3.1 A Declared Synthetic Construction

Methodological Transparency: Read This Section Before Any Number in This Chapter

As stated in the Preface, no raw toponym-by-toponym bigram count data is available within the scope of this volume. Every point cloud analysed below is generated by the following declared procedure, fully reproducible from the script in Appendix 12:

1. Fix a reduced bigram alphabet of size $K = 300$ (illustrative; a real GeoNames bigram alphabet over Latin-script toponyms is larger, but $K = 300$ is large enough that the achievable single-draw entropy range, $[0, \log_2 300] = [0, 8.229]$ bits, comfortably covers the observed range of real zone entropies, $H_2 \in [6.655, 8.174]$ bits, Vol. I Table 4.1).

2. For a target zone with published entropy H_2^{target}, find by bisection the concentration parameter $\alpha > 0$ of a symmetric Dirichlet prior $\text{Dir}(\alpha \mathbf{1}_K)$ such that a single draw $p \sim \text{Dir}(\alpha \mathbf{1}_K)$ has expected Shannon entropy $\mathbb{E}[H(p)] \approx H_2^{\text{target}}$ (estimated by Monte Carlo over 60 draws per bisection step).

3. Draw one such p per zone – this p plays the role of the zone's *population* bigram distribution, analogous to (but not numerically identical with) the $\hat{p}_i$ of *Toponomastic Entropy*.

4. Sample $n = 40$ synthetic toponyms per zone: for each, draw a toponym length $L \sim \text{Poisson}(12)$ (truncated below at 4), then a bigram count vector $\text{Mult}(L, p)$, Laplace-smooth by adding 0.5 to every count, and normalise – giving each synthetic toponym its own empirical distribution $q \in \Delta^{K-1}$, exactly as a real toponym's bigram profile would be estimated.

Table 3.1 reports the calibration achieved for the three representative zones carried through this volume.

Note the monotone relationship between α and entropy: Libya, the least diverse

Table 3.1: Dirichlet calibration for the three representative zones: target entropy H_2^{target} (real, published, Vol. I Table 4.1), the fitted concentration α, and the realised single-draw entropy at convergence. $K = 300$.

Zone	H_2^{target} (bits)	Fitted α	Realised H (bits)
Basque	7.652	1.076	7.637
Armenian	7.540	0.865	7.485
Libya	6.655	0.273	6.788

real zone, requires the most concentrated prior ($\alpha = 0.273$, mass piled onto few bigram types), consistent with *Toponomastic Entropy*, Vol. VII's characterisation of Libyan toponymy as phonotactically narrow; Basque, the most diverse Iberian zone, requires the most diffuse prior ($\alpha = 1.076$). This monotonicity is a sanity check on the calibration procedure, not a new empirical finding – it simply confirms the synthetic generator reproduces the ordering already established by real data.

3.2 Fisher–Rao versus Euclidean Distance on the Same Cloud

For each zone's $n = 40$ synthetic toponyms, two distance matrices were computed on the identical point cloud: the ordinary Euclidean distance $\|q - q'\|_2$ (what linear PCA implicitly uses) and the Fisher–Rao geodesic distance $d_{\text{FR}}(q, q')$ of Proposition 2.2.

Table 3.2: Comparison of Euclidean and Fisher–Rao geometry on the same synthetic 40-toponym cloud, three representative zones. ρ: Pearson correlation between the $\binom{40}{2} = 780$ pairwise Euclidean and Fisher–Rao distances. Mean relative distortion: $\mathbb{E}\big[\,|d_{\text{FR}} - \|\cdot\|_2|/d_{\text{FR}}\,\big]$ over the same pairs. Var. ratio (3 PCs): cumulative variance explained by the first three linear principal components.

Zone	ρ	Mean rel. distortion	Var. ratio (3 PCs)	d_{FR} range (rad)
Basque	0.960	0.923	0.178	[0.261, 0.492]
Armenian	0.970	0.923	0.158	[0.273, 0.474]
Libya	0.962	0.921	0.207	[0.249, 0.486]

Remark: Reading Table 3.2: A Genuine Surprise

The high correlation $\rho \approx 0.96$–0.97 between Euclidean and Fisher–Rao pairwise distances might suggest the two geometries agree closely – but the mean *relative* distortion of ≈ 0.92 tells a different story: on average, the Euclidean distance between two toponyms differs from their true Fisher–Rao geodesic distance by about 92% of the geodesic value itself. The two statistics are not in tension: a high rank correlation is fully consistent with a large, systematic multiplicative distortion, because $\sqrt{q}$ compresses small counts and expands large ones non-linearly relative to raw q (the square-root map of Prop. 2.2 is exactly the correction). The

practical upshot is that Euclidean PCA, applied directly to bigram frequency vectors, preserves the *ordering* of which toponyms are closer to which fairly well, but badly miscalibrates *how much* closer – exactly the boundary-distortion pathology anticipated in Section 1.2, here observed directly rather than merely argued for.

Equally informative is the linear-PCA variance-explained column: even the first three principal components capture only 16–21% of the total variance in the raw bigram vectors for every zone. This is not because the underlying toponym clouds are intrinsically high-dimensional – Proposition 2.2 shows they live, after the isometric square-root embedding, on a sphere whose Fisher–Rao distances span a narrow range (~ 0.25–0.49 radians, Table 3.2, versus a maximum possible $\pi \approx 3.14$) – but because linear PCA is fitting a flat subspace to data whose natural variation is curved and concentrated near a thin spherical shell, a regime in which linear variance decomposition is known to be an inefficient descriptor (Pennec et al., 2020).

Methodological Transparency: What This Section Does and Does Not Show

This is a demonstration of a *generic* geometric phenomenon (linear methods distort distances on the simplex) on *synthetic, calibrated* data, not a finding about real toponymic structure. The magnitude of distortion reported in Table 3.2 depends on the chosen alphabet size $K = 300$, sample length distribution, and Dirichlet concentration, and would differ (likely less severely) for real toponym data with a different effective support size. What is robust, and follows from Proposition 2.2 analytically rather than from this particular simulation, is the qualitative conclusion: the simplex is curved, Fisher–Rao is its intrinsic metric, and Euclidean PCA does not respect that curvature.

Chapter 4

Geodesics and Curvature on the Statistical Manifold

4.1 From a Distance Formula to a Geometry

Proposition 2.2 supplied a closed-form distance, $d_{\mathrm{FR}}(p,q) = 2\arccos\big(\sum_i \sqrt{p_i q_i}\big)$, and Chapter 3 used it as a black box: a number fed to a filtration or a clustering routine. This chapter opens the box. A Riemannian distance is, by definition, the infimum of path lengths between two points, and the path that achieves this infimum – the *geodesic* – carries information the distance alone does not: its existence and uniqueness, its dependence on the metric tensor, and the curvature of the space it traces through, are all properties that matter for any operation more elaborate than "compute one number per pair," such as computing a Fréchet mean (a generalisation of the arithmetic mean to curved spaces) or interpolating between two zones' representative distributions. None of these operations is performed on real data anywhere in this volume; this chapter exists because the synthesis of Chapter 11 commits to a future programme in which they will be, and because the curvature sign derived here explains, retroactively, why the comparison with Euclidean PCA in Chapter 3 works the way it does.

4.2 The Fisher Information Metric in Coordinates

Definition 4.1 (Fisher information metric on the simplex)**.** For $p \in \Delta^{K-1}_{>0}$ (the open simplex, all coordinates strictly positive) with local coordinates $p_1, \dots, p_{K-1}$ ($p_K = 1 - \sum_{i<K} p_i$), the Fisher information metric tensor $\mathcal{I}(p)$ has entries

$$\mathcal{I}_{ij}(p) \;=\; \mathbb{E}_{X\sim p}\!\left[\frac{\partial \log p(X)}{\partial p_i}\,\frac{\partial \log p(X)}{\partial p_j}\right] \;=\; \frac{\delta_{ij}}{p_i} + \frac{1}{p_K}, \qquad i,j = 1,\dots,K-1,$$

where δ_{ij} is the Kronecker delta. The Riemannian length of a tangent vector $v \in T_p\Delta^{K-1}$ is then $\|v\|_p^2 = v^\top \mathcal{I}(p)\, v$, and the length of a path $\gamma : [0,1] \to \Delta^{K-1}$ is $L(\gamma) = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, dt$.

Derivation of the coordinate formula. For a single categorical draw with parameter p, $\log p(X{=}i) = \log p_i$ for $i < K$ and $\log p(X{=}K) = \log(1 - \sum_{i<K} p_i)$. Then

$\partial \log p(X{=}i)/\partial p_j = \delta_{ij}/p_i$ for $i < K$, and $\partial \log p(X{=}K)/\partial p_j = -1/p_K$ for all $j < K$. Taking the expectation over X, the cross term contributes $\Pr(X{=}K) \cdot (1/p_K)(1/p_K) = 1/p_K$ to every entry, while the $i = j < K$ diagonal term additionally receives $\Pr(X{=}i) \cdot (1/p_i)^2 = 1/p_i$ from the $X{=}i$ outcome; off-diagonal $i \neq j$, $i, j < K$ entries receive no contribution from $X{=}i$ or $X{=}j$ since the two partial derivatives involve disjoint indicator supports, leaving only the $1/p_K$ cross term. Collecting terms gives the stated formula. □

Proposition 4.2 (The square-root map is an isometry onto a sphere)**.** *Let $\phi : \Delta^{K-1}_{>0} \to \mathbb{S}^{K-1}_{>0}$, $\phi(p) = (\sqrt{p_1}, \ldots, \sqrt{p_K})$, where $\mathbb{S}^{K-1}_{>0}$ is the positive orthant of the unit sphere of radius 1 in $\mathbb{R}^K$ under the ordinary Euclidean (round) metric, scaled by a factor of 4. Then ϕ is a diffeomorphism onto its image, and pulling back the round metric of radius $\frac{1}{2}$ along ϕ recovers $\mathcal{I}$ exactly: $\phi^*(g_{\text{round},1/2}) = \mathcal{I}$. Consequently the Fisher–Rao geodesic distance between $p, q \in \Delta^{K-1}_{>0}$ equals 1 (radius $\frac{1}{2}$ great-circle arc length) times the angle between $\phi(p)$ and $\phi(q)$, i.e. $d_{\text{FR}}(p, q) = 2 \arccos\big(\phi(p) \cdot \phi(q)\big) = 2 \arccos\big(\sum_i \sqrt{p_i q_i}\big)$, recovering Proposition 2.2.*

Proof. Write $u_i = \sqrt{p_i}$, so $p_i = u_i^2$ and $\sum_i u_i^2 = 1$: the image of ϕ lies on the unit sphere $\mathbb{S}^{K-1}$ by construction. Differentiating, $du_i = \frac{1}{2\sqrt{p_i}}\, dp_i$, so the round metric pulled back is $\sum_i du_i^2 = \sum_i \frac{1}{4p_i}\, dp_i^2$ on the chart where all K coordinates are free subject to $\sum p_i = 1$; restricting to the $(K{-}1)$ independent coordinates $p_1, \ldots, p_{K-1}$ (with $p_K = 1 - \sum_{i<K} p_i$, so $dp_K = -\sum_{i<K} dp_i$) and expanding the quadratic form reproduces, after collecting cross terms exactly as in the proof of Definition 4.1, the matrix $\delta_{ij}/p_i + 1/p_K$ up to the overall factor of 4 from the $\frac{1}{4p_i}$ coefficient, which is absorbed by taking the round metric of radius $\frac{1}{2}$ rather than 1 (geodesic distances on a sphere of radius r scale as r times the angle, and $r = \frac{1}{2}$ combined with the angle $\arccos(\cdot)$ gives the factor of 2 in $d_{\text{FR}} = 2\arccos(\cdot)$). Geodesic distance on a sphere under the round metric is the great-circle arc length, $r \arccos(\hat{u} \cdot \hat{v})$ for unit vectors $\hat{u}, \hat{v}$; substituting $r = \frac{1}{2}$ and $\hat{u} \cdot \hat{v} = \sum_i \sqrt{p_i q_i}$ (the Hellinger affinity, this volume's $\sum_i u_i v_i$) gives the stated distance formula. □

4.3 Curvature: Why the Simplex Is Not Flat

Proposition 4.3 (Constant positive sectional curvature)**.** *Under the Fisher information metric, $\Delta^{K-1}_{>0}$ has constant sectional curvature $\kappa = 4$ at every point and in every two-plane, inherited from the unit-radius round sphere of Proposition 4.2 (radius $\frac{1}{2}$, whose constant sectional curvature is $1/r^2 = 4$).*

Proof. The round sphere of radius r has constant sectional curvature $1/r^2$ at every point, a standard fact of Riemannian geometry (the sphere is a space form). Proposition 4.2 exhibits $(\Delta^{K-1}_{>0}, \mathcal{I})$ as isometric to an open subset (the positive orthant) of the round sphere of radius $\frac{1}{2}$. Sectional curvature is a local isometry invariant, so it transfers unchanged: $\kappa \equiv 1/(\frac{1}{2})^2 = 4$ on the simplex. □

Remark 4.4 (What positive curvature implies for the comparisons of Chapter 3). A space of constant positive curvature has two properties directly relevant to this volume's empirical comparisons. First, geodesic triangles are "fatter" than their Euclidean counterparts: the sum of interior angles exceeds π, and consequently chord

(straight-line, Euclidean) distances systematically *underestimate* true geodesic distance, with the gap growing with the angular separation – this is the quantitative form of Chapter 3's qualitative claim that flattening the simplex by ordinary PCA distorts large pairwise distances more than small ones. Second, the exponential map $\exp_p : T_p\Delta^{K-1} \to \Delta^{K-1}$ (which sends a tangent vector v to the point reached by following the geodesic in direction v for unit time) is well-defined and a local diffeomorphism for $\|v\| < \pi/\sqrt{4} = \pi/2$ but develops conjugate points beyond that radius – a geometric obstruction, not merely a numerical one, to defining geodesic interpolation between two distributions whose Fisher–Rao distance already exceeds $\pi/2$. Restated for this series: the largest pairwise Fisher–Rao distance observed in Table 7.1-scale data (cutoffs around 0.48–0.50 radians, well under $\pi/2 \approx 1.571$) is comfortably inside the conjugate-point-free regime, so this concern is not live for the present volume's specific numbers, but it would become relevant for any future application to a corpus whose entropy range is wide enough to push pairwise Fisher–Rao distances past $\pi/2$.

4.4 The Fréchet Mean: Algorithm and an Explicit Computation

Definition 4.5 (Fréchet mean)**.** For a point cloud $p^{(1)}, \dots, p^{(n)} \in \Delta^{K-1}_{>0}$, the Fréchet mean under d_{FR} is $\bar{p} = \arg\min_{q \in \Delta^{K-1}} \sum_{j=1}^{n} d_{\mathrm{FR}}(q, p^{(j)})^2$, the curved-space generalisation of the Euclidean arithmetic mean as a sum-of-squared-distances minimiser.

On the round sphere of Proposition 4.2, the Fréchet mean of a point cloud with diameter less than $\pi/2$ (Karcher's theorem) exists and is unique, and can be computed by a standard iterative algorithm: at the current estimate μ, map every cloud point x_j into the tangent space $T_\mu\mathbb{S}^{K-1}$ via the spherical logarithm $\log_\mu(x) = \theta\,\dfrac{x - \cos\theta\,\mu}{\sin\theta}$, $\theta = \arccos(\mu \cdot x)$; average the resulting tangent vectors $\bar{v} = \frac{1}{n}\sum_j \log_\mu(x_j)$; and move μ along the geodesic in direction $\bar{v}$ via the spherical exponential $\exp_\mu(\bar{v}) = \cos\|\bar{v}\|\,\mu + \sin\|\bar{v}\|\,\bar{v}/\|\bar{v}\|$, iterating to convergence ($\|\bar{v}\| < 10^{-10}$). Every per-zone Fisher–Rao cloud analysed in Chapter 7 satisfies the diameter condition (Table 7.1's cutoffs are all under $\pi/2$), so this algorithm is guaranteed to converge to a well-defined mean for each of the three representative zones.

Remark: This Algorithm Has Now Been Run

Unlike in an earlier draft of this chapter, the iterative algorithm above was executed (Appendix 12, extended script) on a fresh draw of the three zones' synthetic clouds, generated by the identical procedure of Section 3.1 (same target entropies, same $K = 300$, same $n = 40$, seed fixed for reproducibility). The resulting calibration α values (Table 4.1) differ slightly from Table 3.1 because the Monte Carlo bisection and the multinomial sampling both consume fresh random draws on a fresh run; this is expected stochastic variation in a declared synthetic procedure, not an inconsistency, and the qualitative ordering (Libya most concentrated, Basque least) is preserved.

Table 4.1: Fréchet mean computation on a fresh draw of the three representative synthetic clouds ($n = 40$, $K = 300$). SSE (Fréchet) and SSE (Euclidean centroid) are $\sum_j d_{\mathrm{FR}}(\bar{p}, x_j)^2$ for the Fréchet mean and for the ordinary coordinatewise (Euclidean) centroid respectively, both measured in the *same* Fisher–Rao metric. d_{FR} gap is the Fisher–Rao distance between the two candidate centres. Iterations: number of fixed-point updates to convergence.

Zone	Fitted α	Iter.	SSE (Fréchet)	SSE (Eucl. centroid)	d_{FR} gap (rad)
Basque	1.055	4	3.080	3.097	0.0205
Armenian	0.851	4	3.004	3.022	0.0217
Libya	0.272	4	2.776	2.806	0.0271

Remark: Reading Table 4.1: A Small but Consistent Gap, and a More Interesting Artefact

Two findings emerge. First, the gap itself: the Fréchet mean achieves strictly lower Fisher–Rao sum-of-squared-distances than the Euclidean centroid in every zone, as the theory guarantees (the Fréchet mean is the unique *minimiser* of that quantity), but the margin is small, 1–2% of the SSE itself, and shrinks further as the concentration α shrinks (Libya). This is not surprising in retrospect: convergence took only 4 iterations in every zone, and the Fisher–Rao gap between the two candidate centres (0.02–0.03 radians) is an order of magnitude smaller than the cloud's own spread (mean Fisher–Rao distance to either centre, ≈ 0.26–0.28 radians) – consistent with the curvature discussion of Section 4.4: at distances this far inside the conjugate-point radius $\pi/2 \approx 1.571$, the sphere's curvature has not yet had room to make the Euclidean chord-based centroid a badly miscalibrated stand-in for the geodesic-based mean.
Second, and unanticipated: both candidate centres have realised Shannon entropy, $H(\bar{p}) \approx 8.22$ bits, indistinguishable from the maximum possible $\log_2 300 = 8.229$ bits – regardless of which zone, even Libya, whose *individual* target entropy is 6.655 bits. The explanation is a genuine artefact of the synthetic construction's Laplace smoothing interacting with sparsity: each synthetic toponym has length $L \sim 12$, so its raw multinomial draw populates at most 12 of the $K = 300$ bigram bins, and the remaining ≥ 288 bins each receive only the Laplace floor of 0.5; once normalised, a single toponym's distribution is therefore dominated by this near-uniform floor rather than by the handful of bins it actually sampled, and *any* centre computed by averaging 40 such floor-dominated distributions – Euclidean or Fréchet, it makes no difference – inherits the same near-uniform character. This is a property of the per-toponym smoothing choice ($+0.5$ to all K bins regardless of $L \ll K$), not of the geometry being compared, and it is recorded here as a caution for any future volume that estimates a *centre* (rather than pairwise distances, which are far less sensitive to the floor) from individually-smoothed sparse count vectors: the floor should scale with, or be compared against, the ratio L/K, not fixed at a constant additive value.

Corollary 4.6. *The Fréchet mean under d_{FR} of a Dirichlet$(\alpha \mathbf{1}_K)$ sample, for sample sizes $n \to \infty$, converges almost surely to ϕ^{-1} applied to the Euclidean mean of $\phi(p^{(j)})$ on the sphere, which by symmetry of the Dirichlet distribution under coordinate permutation is*

$\phi^{-1}(c \cdot \mathbf{1}_K/\sqrt{K})$ for a normalising constant c, i.e. the uniform distribution $\mathbf{1}_K/K$ – recovering, asymptotically, the generating distribution's own mean parameter direction. The finite-n computation of Table 4.1 is consistent with, but should not be conflated with, this asymptotic statement: at $n = 40$ the entropy saturation discussed above is driven by per-toponym smoothing, a finite-L effect entirely distinct from the $n \to \infty$ averaging limit described by this corollary, and the two effects happen to push the realised mean in the same direction (toward the uniform distribution) for unrelated reasons.

4.5 The Geodesic Equation and Explicit Interpolation via Spherical Linear Interpolation

Proposition 4.3 and the remark following it establish *that* the simplex under $\mathcal{I}$ has constant positive curvature and *where* the exponential map stops being a diffeomorphism, but neither result exhibits a geodesic explicitly. This section closes that gap: it derives the actual geodesic path between two points of $\Delta^{K-1}_{>0}$, in closed form, by transporting the classical formula for a great-circle arc back through the isometry ϕ of Proposition 4.2.

Definition: Geodesic

A curve $\gamma : [0,1] \to M$ on a Riemannian manifold (M, g) is a geodesic if it is a critical point of the length functional $L(\gamma) = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, dt$ among curves with fixed endpoints, equivalently (in any local coordinate chart) if it satisfies the geodesic equation

$$\ddot\gamma^k(t) \;+\; \sum_{i,j} \Gamma^k_{ij}(\gamma(t))\, \dot\gamma^i(t)\, \dot\gamma^j(t) \;=\; 0,$$

where $\Gamma^k_{ij} = \frac{1}{2} g^{kl}\big(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\big)$ are the Christoffel symbols of g.

Rather than solve this second-order system directly in the $p_1, \ldots, p_{K-1}$ coordinates of Definition 4.1 – where the Christoffel symbols of $\mathcal{I}$ are algebraically unpleasant, on account of the $1/p_K$ cross term – this section exploits Proposition 4.2 directly: since ϕ is an isometry onto (an open subset of) a round sphere, and isometries carry geodesics to geodesics, the geodesics of $\Delta^{K-1}_{>0}$ are exactly the ϕ-preimages of the geodesics of the sphere, and the latter are classical.

Lemma 4.7 (Geodesics of the round sphere are great circles)**.** *On the unit sphere $\mathbb{S}^{K-1} \subset \mathbb{R}^K$ with the round metric, a curve is a geodesic if and only if it lies on a great circle, i.e. the intersection of $\mathbb{S}^{K-1}$ with a 2-plane through the origin.*

Proof. A great circle through unit vectors u, v (with $u \cdot v \neq \pm 1$) can be parametrised as $c(t) = \cos(t)\, u + \sin(t)\, \hat w$, where $\hat w$ is the unit vector in $\mathrm{span}(u, v)$ orthogonal to u with $v \cdot \hat w > 0$; this curve satisfies $c(t) \cdot c(t) = 1$ for all t by construction, so it lies on the sphere, and $\ddot c(t) = -c(t)$, which is everywhere normal to the sphere (parallel to the position vector $c(t)$ itself, the sphere's outward normal at that point). A curve's acceleration being purely normal to the manifold it lies on is exactly the coordinate-free characterisation of a geodesic (the tangential component of acceleration, which is what the Christoffel-symbol term in the geodesic equation measures and removes, vanishes identically here), so every great circle is a geodesic.

Conversely, the sphere's geodesic equation is a second-order ODE, so by the Picard–Lindelöf existence and uniqueness theorem, for any starting point and initial velocity there is a unique maximal geodesic with that initial data; the great circle through that point with that initial tangent direction already satisfies the initial data and is a geodesic by the argument above, so by uniqueness it *is* the geodesic. □

Proposition 4.8 (Explicit geodesics on the simplex via slerp)**.** *Let $p, q \in \Delta_{>0}^{K-1}$ with $d_{\mathrm{FR}}(p, q) < \pi$ (equivalently, $p \neq q$ are not antipodal under ϕ, which cannot happen for two probability vectors in the open simplex in any case), and set $\theta = \frac{1}{2} d_{\mathrm{FR}}(p, q) = \arccos\big(\sum_i \sqrt{p_i q_i}\big)$. Then the unique unit-speed geodesic from p to q, reparametrised to $\gamma : [0, 1] \to \Delta_{>0}^{K-1}$ with $\gamma(0) = p$, $\gamma(1) = q$, is given coordinatewise by*

$$\gamma(t)_i \;=\; \left(\frac{\sin\big((1-t)\theta\big)}{\sin\theta} \sqrt{p_i} \;+\; \frac{\sin(t\theta)}{\sin\theta} \sqrt{q_i} \right)^2, \qquad i = 1, \ldots, K.$$

Proof. Write $u = \phi(p) = (\sqrt{p_1}, \ldots, \sqrt{p_K})$ and $v = \phi(q) = (\sqrt{q_1}, \ldots, \sqrt{q_K})$, both unit vectors in $\mathbb{R}^K$ with $u \cdot v = \cos\theta$ by Proposition 4.2. The great circle through u and v, parametrised proportionally to arc length over $t \in [0, 1]$ so that it runs from u at $t = 0$ to v at $t = 1$, is the standard *spherical linear interpolation* (slerp) curve

$$w(t) \;=\; \frac{\sin\big((1-t)\theta\big)}{\sin\theta}\, u \;+\; \frac{\sin(t\theta)}{\sin\theta}\, v,$$

a curve that is easily checked to lie on $\mathbb{S}^{K-1}$ (a direct trigonometric expansion gives $\|w(t)\|^2 = 1$ for all t, using $u \cdot u = v \cdot v = 1$ and $u \cdot v = \cos\theta$) and to pass through u, v at $t = 0, 1$ respectively; by Lemma 4.7 it is therefore the geodesic between them. By Proposition 4.2, ϕ is an isometry from $(\Delta_{>0}^{K-1}, \mathcal{I})$ onto its image, and an isometry maps geodesics to geodesics with matching endpoints; hence $\gamma = \phi^{-1} \circ w$ is the geodesic on the simplex from p to q. Since $\phi^{-1}(w)_i = w_i^2$ coordinatewise (inverting the square root that defines ϕ), applying this to $w(t)$ gives the stated formula. □

Corollary 4.9 (The naive Euclidean interpolant is not the geodesic)**.** *The coordinatewise linear interpolant $\tilde{\gamma}(t) = (1-t)p + tq$ (renormalised to sum to 1, since $(1-t)p + tq$ already does whenever p, q do) coincides with $\gamma(t)$ of Proposition 4.8 only in the degenerate limit $\theta \to 0$ ($p = q$); for any $\theta > 0$ the two curves differ at every interior $t \in (0, 1)$, since γ traces the constant-curvature-4 great circle of Lemma 4.7 while $\tilde{\gamma}$ traces a Euclidean chord, and a great circle coincides with its chord only at its two endpoints.*

Remark: Why This Formula Is Recorded Here, and Not Used Anywhere Else in This Volume

Proposition 4.8 answers, explicitly, a question Section 4.4's iterative Fréchet-mean algorithm only needed implicitly (that algorithm moves *along* a geodesic at each update step, via $\exp_\mu(\bar{v})$, without ever writing the full path down). No chapter of this volume interpolates between two zones' representative distributions – the operation Section 4.1 flagged as a "future programme," not a present one – so Proposition 4.8 is not applied to any number reported elsewhere in this book. It is included because Corollary 4.9 makes concrete, for the first time in this series, exactly what "curved" means operationally: a modeller who interpolated between two zones' phonotactic profiles by linear mixture of their raw bigram distribu-

tions – a natural first instinct – would not be tracing the Fisher–Rao geodesic at all, but a Euclidean chord that cuts inside the sphere and back out, momentarily passing through distributions with *lower* total Fisher information distance to both endpoints combined than the true geodesic requires; the explicit formula above is what a future volume would need to interpolate correctly instead.

Chapter 5

Literature Review: Information Geometry in Linguistics, Phylogenetics, and Genomics

5.1 Why a Literature Review Belongs Here, Not in the Preface

The Preface stated this volume's relation to its own series; this chapter situates it within the wider use of information-geometric and topological methods on biological and linguistic data, a literature that is older, larger, and more methodologically varied than a single footnote can convey. The aim is not completeness but orientation: a reader who has followed Chapters 2–3 should leave this chapter knowing which parts of the present construction are well-trodden ground, which are genuine adaptations, and which remain open even in fields with far larger corpora than historical toponymy.

5.2 The Fisher–Rao Metric Outside Information Geometry Proper

Čencov's uniqueness theorem (Čencov, 1982) – that the Fisher–Rao metric is, up to scale, the only Riemannian metric on a statistical manifold invariant under sufficient statistics – is a purely mathematical-statistical result, proved without reference to any particular application domain. Its first systematic exposition as a general framework is Amari and Nagaoka (2000), whose treatment of dual affine connections (e-connection and m-connection) and the associated Bregman divergences this volume has deliberately not invoked, restricting itself to the Riemannian (Fisher–Rao) structure alone; Chapter 3's remark that the present volume is a restriction of the full information-geometric apparatus to its metric component is a direct consequence of this choice, made for tractability rather than because the dual-connection structure is inapplicable to compositional toponymic data.

Compositional data analysis. A bigram-frequency vector is, in the vocabulary of Aitchison (1986), a *composition*: a vector of strictly positive parts summing to a con-

stant, for which ordinary Euclidean statistics (means, PCA, regression) are known to be inappropriate because the sample space is the simplex, not $\mathbb{R}^K$. Aitchison's own resolution – the log-ratio transformation, mapping Δ^{K-1} to $\mathbb{R}^{K-1}$ via $x \mapsto \log(x_i/x_K)$ – is a different geometric move from the present volume's: it flattens the simplex into Euclidean space by an explicit coordinate change, rather than retaining its curvature and working with geodesics directly. Both approaches address the same underlying problem (Euclidean PCA on raw proportions is geometrically incoherent), and the choice between them is not settled by mathematics alone: log-ratio coordinates make standard multivariate machinery (linear regression, covariance estimation) directly applicable at the cost of requiring every component to be strictly positive (a zero count forces an arbitrary pseudocount, exactly the Laplace-smoothing step this series already performs for unrelated reasons in *Toponomastic Entropy*), while the Fisher–Rao route of this volume requires no such modification but sacrifices linearity, making every downstream computation (means, regression, clustering) a genuinely Riemannian rather than Euclidean problem. A systematic comparison of the two – log-ratio PCA versus Fisher–Rao geodesic PCA on the same real toponym-level data – is listed as an open methodological question in Chapter 11 and remains undone here for the same data-availability reason that constrains the rest of this volume.

Phylogenetics and population genetics. Allele-frequency vectors at a locus with K alleles are, formally, points on Δ^{K-1}, and Fisher's original 1922 motivation for the metric that bears his name was genetic: the rate of genetic drift under the Wright–Fisher model is governed by exactly the arccos-geodesic distance of Proposition 2.2, with 2 replaced by a population-size-dependent diffusion constant. The square-root transform of Proposition 2.2, mapping allele frequencies onto the positive orthant of a sphere, is the same "Fisher–Rao = arccos on the sphere" identity used throughout population genetics under the name *Cavalli-Sforza chord distance* (Pennec et al., 2020), independently rediscovered in that literature decades before its formal information-geometric statement. This is a substantive precedent for the present volume's central identity, not a coincidence: any field whose primitive observation is a multinomial count vector over a fixed category set – alleles, bigrams, word types – inherits the same simplex geometry, and the chord-distance identity of Proposition 2.2 is exactly the move that lets genetics treat allele-frequency change as motion on a sphere rather than an arbitrary high-dimensional space. The disanalogy is equally instructive: population genetics has, for many loci, direct access to individual-level genotypes (the "toponym-level" data this volume lacks), so the corresponding Riemannian-statistics machinery there is typically applied to real rather than calibrated-synthetic point clouds, a data-availability gap that is this volume's defining limitation rather than the method's.

Linguistics. Direct prior use of Riemannian information geometry on linguistic corpora is comparatively rare. The closest precedents are in computational phonology and corpus linguistics, where KL divergence between n-gram or phoneme-frequency distributions is used as a (non-geodesic, asymmetric) distance for dialect or language clustering – precisely the proxy-distance approach *Toponomastic Networks*, Vol. I and *Toponomastic Entropy* both use, and precisely what Chapter 2's remark on KL's failure to be a metric was written to motivate replacing. The present volume's use of the symmetrised, metric Fisher–Rao distance in place of asymmetric KL divergence is, so far as a literature search internal to this series can determine, a novel application within toponymy specifically, though the underlying metric is, as

the genetics precedent above shows, far from new in adjacent fields working with the same mathematical object (a frequency vector over a fixed finite alphabet).

Hyperbolic and Fisher-geometric embeddings in NLP. A more recent strand of work, entirely posterior to the genetics and compositional-data precedents above, applies non-Euclidean geometry to learned rather than counted linguistic representations. Nickel and Kiela (2017) embed hierarchically structured symbolic data (WordNet's hyponymy tree, among other examples) into the Poincaré ball rather than Euclidean space, on the grounds that hyperbolic geometry's exponential volume growth accommodates tree-like hierarchy with far lower distortion than a flat embedding of comparable dimension can achieve. Tifrea et al. (2019) extend this to distributional word embeddings proper, training a hyperbolic analogue of GloVe and showing that the Fisher information metric of a Gaussian distribution – the same object Chapter 2 restricts to the categorical (simplex) case – is itself hyperbolic rather than spherical, a fact traced to Costa et al. (2015) and connected there to the Gaussian-embedding proposal of Vilnis and McCallum (2015). This is a genuinely informative contrast with the present volume's construction rather than a further precedent for it: a bigram-frequency vector is a point on the *categorical* simplex, whose Fisher–Rao metric Proposition 2.2 identifies with a *positively curved, spherical* geometry (an arccos distance, Proposition 2.2), whereas a Gaussian's Fisher metric is *negatively curved, hyperbolic*. Both literatures start from Čencov's uniqueness theorem and the same Fisher-information construction, and both replace a naive Euclidean comparison with a curved one, but they arrive at opposite-signed curvature because the underlying statistical family differs (categorical versus Gaussian) – a reminder that "information geometry" is not a single curvature prescription but a construction whose sign and shape are dictated by which exponential family is in play, and that the appropriateness of the spherical route taken here rests specifically on toponymic bigram counts being categorical data, not on any general argument that curved geometry is preferable to flat.

5.3 Persistent Homology Outside Pure Topology

Edelsbrunner and Harer (2010) and Zomorodian and Carlsson (2005) established persistent homology as a computational tool rather than a purely theoretical construction; its subsequent uptake outside pure mathematics has been driven almost entirely by applications to high-dimensional point clouds where the "shape" of the data – connectedness, holes, voids – is itself the scientific question, rather than a nuisance to be averaged away. Three application domains are worth contrasting with the present one.

Genomics and single-cell biology. Persistent homology applied to gene-expression point clouds (cells as points in expression space) has detected genuinely non-trivial H_1 structure – loops corresponding to cyclic processes such as the cell cycle – that linear methods (PCA, hierarchical clustering) miss by construction, because a cycle has no privileged "centre" for a linear method to anchor on. This is the single most direct precedent for why Chapter 7 treats an H_1 loop, were one to appear robustly, as a qualitatively different finding from anything Euclidean PCA could report: the genomics literature's loops are real, biological, and would have been invisible to any of this series' existing entropy-based or graph-based methods.

Materials science and sensor networks. Persistent homology of point clouds sampled from physical structures (porous materials, sensor coverage regions) is used to certify coverage gaps (H_1/H_2 holes) with formal stability guarantees under measurement noise – the practical motivation behind the stability theorems of Chapter 10, since a method that is not provably stable under small perturbations of the input distance matrix would be unusable for noisy sensor data, and the same stability property is what licenses, in the present volume, treating short-lived H_1 bars as artefacts rather than weak signal.

Natural language processing. A smaller body of work applies persistent homology directly to word- or document-embedding point clouds (rather than to a derived distance structure such as Fisher–Rao), typically using ordinary Euclidean distance in the embedding space. This is the closest existing precedent to Chapter 7's pipeline in domain (language) but the furthest in geometric assumption (Euclidean rather than Fisher–Rao); the present volume's combination – TDA on an information-geometric rather than Euclidean distance matrix, applied to toponymic rather than general lexical data – does not appear to have a direct precedent in either the information-geometry or the computational-linguistics literature consulted in preparing this volume, which is the methodological contribution this part of the tetralogy claims, subject to the data-availability caveat restated throughout.

5.4 What This Comparison Does and Does Not License

Remark: The Limits of a Literature Review

None of the parallels drawn in this chapter substitute for direct validation on real toponymic data. That genetics successfully applies Fisher–Rao geometry to real allele-frequency point clouds, or that genomics successfully detects real H_1 loops in real expression data, shows that the mathematical machinery of Chapters 2–7 is not a methodological dead end – it has worked, on comparably structured data, elsewhere. It does not show, and is not offered as evidence, that real toponym-level bigram clouds have any particular geometric or topological structure; that question remains exactly as open as the Preface and Section 7.2 state it.

Chapter 6

Topological Data Analysis: Persistent Homology

6.1 From a Point Cloud to a Filtration of Simplicial Complexes

Definition: Vietoris–Rips Complex

Given a finite metric space (X, d) and a scale parameter $\epsilon \geq 0$, the Vietoris–Rips complex $\mathrm{VR}_\epsilon(X)$ is the simplicial complex whose k-simplices are all $(k+1)$-subsets of X with pairwise distance at most ϵ. As ϵ increases from 0 to $\mathrm{diam}(X)$, $\mathrm{VR}_\epsilon(X)$ grows monotonically (every simplex present at scale ϵ remains present at scale $\epsilon' > \epsilon$), defining a *filtration*.

Definition: Persistent Homology

For each homological dimension p (here $p = 0$: connected components; $p = 1$: independent loops), the persistent homology of the filtration tracks, for every topological feature that is born at some scale ϵ_{birth} (first appears) and dies at ϵ_{death} (is filled in or merges with an older feature), the interval $[\epsilon_{\text{birth}}, \epsilon_{\text{death}})$, called a *bar*. The collection of bars across all features and dimensions is the *persistence barcode*; features with long bars (large $\epsilon_{\text{death}} - \epsilon_{\text{birth}}$) are interpreted as robust topological structure, and short bars as noise.

For dimension 0, persistent homology has an exact, classical description: $\mathrm{VR}_\epsilon(X)$'s connected components, as ϵ grows, merge exactly as in single-linkage hierarchical clustering, so the H_0 barcode is precisely the single-linkage dendrogram, with one bar $[0, \epsilon)$ per merge event plus one infinite bar for the final, fully-connected component. For dimension 1 (loops), no such shortcut exists in general; the standard algorithm reduces the filtration's *boundary matrix* (over $\mathbb{F}_2$, the field with two elements) by a Gaussian-elimination-like procedure, matching each simplex that creates a cycle to the simplex that later fills it in (Edelsbrunner and Harer, 2010; Zomorodian and Carlsson, 2005).

Algorithm 1 Boundary-matrix reduction (standard persistence algorithm)

Sort all simplices by filtration value, ties broken by dimension
For each simplex σ, represent its boundary $\partial\sigma$ as a column (set of simplex-indices) over $\mathbb{F}_2$
for c in increasing filtration order **do**
    while column c is non-empty **and** its lowest non-zero index ℓ already a pivot of an earlier column c' **do**
        column $c \leftarrow$ column c XOR column c'
    end while
    if column c is non-empty **then**
        register $\ell = \mathrm{low}(c)$ as a pivot of column c
        emit a bar: birth at simplex ℓ, death at simplex c
    else
        simplex c remains an unmatched birth (essential class, or later paired)
    end if
end for

Remark: An Honest Implementation Note

The TDA pipeline of this volume was implemented directly (boundary-matrix reduction over $\mathbb{F}_2$, restricted to dimensions 0 and 1, full filtration range), rather than relying on a third-party persistent homology library, none of which was available in the present computing environment. The implementation (Appendix 12) is exact, not approximate, for the point-cloud sizes used here ($n = 40$, giving $\binom{40}{2} = 780$ candidate edges and up to $\binom{40}{3} = 9880$ candidate triangles before filtration restricts both sets); for larger point clouds, the triangle count would require either a sparse (witness-complex) approximation or a dedicated library, out of scope here.

6.2 A Worked Toy Example

Before applying Algorithm 1 to the 40-point synthetic toponym clouds of Chapter 7, it is worth tracing it by hand on a point set small enough to verify every step: six points placed at the vertices of a regular hexagon of unit side, plus one outlier point well separated from the hexagon (ordinary Euclidean distance in $\mathbb{R}^2$, *not* Fisher–Rao – this section is pure geometry, illustrating the algorithm, not toponymic content).

Table 6.1: The three distinct edge lengths of the unit hexagon-plus-outlier point set, and the filtration values at which the relevant topological events occur.

Edge type	**Length**	**Event**
Hexagon side (6 edges)	1	loop is born (H_1)
Short diagonal, skip-1 (6 edges)	$\sqrt{3} \approx 1.732$	loop dies (H_1); hole triangulated
Long diagonal, skip-2 (3 edges)	2	(irrelevant: hole already filled)
Outlier to nearest hexagon vertex	≈ 3.287	outlier joins main component (H_0)

Tracing Algorithm 1: for $\epsilon < 1$, all 7 points are isolated, so H_0 has 7 live bars

and H_1 is empty. At $\epsilon = 1$, all 6 hexagon-perimeter edges enter the filtration simultaneously, merging the 6 hexagon vertices into a single component (five merge events, each killing one finite H_0 bar of length $1-0 = 1$) and, because the 6 edges close into a cycle with no 2-simplex yet present to fill it, giving birth to one H_1 class: the hexagon's perimeter loop. The outlier remains a separate component, its H_0 bar still open. Nothing changes until $\epsilon = \sqrt{3} \approx 1.732$, when all 6 short diagonals enter simultaneously, together with the 6 triangles they complete with adjacent perimeter edges (e.g. vertices $0, 1, 2$: edges $(0, 1)$ and $(1, 2)$ have length 1, diagonal $(0, 2)$ has length $\sqrt{3}$, so the triangle $\{0, 1, 2\}$ enters at $\epsilon = \sqrt{3}$). These six triangles, together with the three "long" triangles such as $\{1, 3, 5\}$ (every pairwise distance among $1, 3, 5$ is exactly $\sqrt{3}$, also entering at this filtration value), fan-triangulate the hexagon's interior completely, so the perimeter loop becomes the boundary of a filled 2-chain and the H_1 class dies. Its bar is therefore $[1, \sqrt{3})$, lifetime $\sqrt{3} - 1 \approx 0.732$. The outlier's H_0 bar remains open until $\epsilon \approx 3.287$ (its distance to the nearest hexagon vertex), at which point the last finite H_0 bar closes, length $\approx 3.287 - 0 = 3.287$, and one essential (infinite) H_0 bar remains, as required for a 7-point filtration run to full connectivity ($n - 1 = 6$ finite H_0 bars plus 1 essential class).

Remark: Why This Example Was Chosen

The hexagon-plus-outlier construction isolates, in a single small example, both signal types Section 6.2 distinguishes. The H_1 bar $[1, \sqrt{3})$ has lifetime ≈ 0.732, which is 73% of its own birth scale – a genuinely long-lived loop, the topological signature a real cyclic arrangement of points would produce, and the kind of bar that did *not* appear (Table 7.1) when the same algorithm was run on synthetic, unimodally-generated toponym clouds. The outlier's H_0 bar, $[0, 3.287)$, is long relative to the other six finite H_0 bars (all length 1, from the simultaneous hexagon closure) precisely because it is genuinely separated from the rest of the cloud, illustrating the H_0 interpretation of Section 6.2: a long-lived H_0 bar flags a point, or small group of points, that stays distinct from the rest of the cloud even under a generous similarity threshold. Both features are, by construction, exactly the kind of structure the real toponym clouds of Chapter 7 did not exhibit – making this toy example a calibration of what the pipeline *would* have reported, had such structure been present.

In the present setting, an H_0 bar that survives to a large ϵ relative to its neighbours signals a toponym, or small group of toponyms, that remains distinct from the rest of the zone's cloud even under a generous similarity threshold – a candidate for a phonotactically unusual sub-population within the zone (an internal-substrate signal that no zone-level entropy statistic can see, cf. Question 1 of Section 1.2). An H_1 bar that survives to a large ϵ signals a *loop*: a cyclic arrangement of toponyms not explainable as variation around a single centre, which would be consistent with (for instance) a zone whose toponyms alternate between two phonotactic regimes without a continuous gradient between them. Short-lived bars of either dimension, by contrast, are the topological analogue of sampling noise, and are not interpreted as structure.

Chapter 7

Applying the Pipeline to the Three Representative Zones

7.1 Persistence Results

Methodological Transparency: Synthetic Data, Restated

As throughout this volume, the point clouds analysed here are the calibrated synthetic constructions of Section 3.1 ($n = 40$ toponyms per zone, Dirichlet–multinomial generative model fitted to match each zone's real published entropy). The persistent-homology *pipeline* below is exact and general; the specific barcodes reported are properties of this synthetic null model, not measurements of real toponymic shape.

For each zone, the Vietoris–Rips filtration was built on the Fisher–Rao distance matrix d_{FR} of Section 3.2, over the full observed range of pairwise distances (cutoff ≈ 0.48–0.50 radians, the per-zone maximum), and reduced via Algorithm 1 in dimensions 0 and 1.

Table 7.1: Persistent homology of the three representative synthetic toponym clouds ($n = 40$ each), full-range Vietoris–Rips filtration on Fisher–Rao distance. H_0: connected-component bars (always $n - 1 = 39$ finite $+ 1$ essential, since the filtration is run to the point of full connectivity). H_1: independent-loop bars; "max lifetime" is the longest birth-to-death interval observed; "lifetime / scale" divides this by the filtration cutoff, giving a scale-free measure of how persistent the most persistent loop is.

Zone	Cutoff (rad)	H_1 bars	Max lifetime	Lifetime/scale
Basque	0.497	4	0.0050	1.0%
Armenian	0.478	16	0.0110	2.3%
Libya	0.491	13	0.0150	3.1%

Remark: Interpreting an Absence of Structure

Every H_1 bar across all three zones is short-lived: the longest-lived loop in any zone persists for at most 3.1% of the filtration's total scale (Libya), and most loops persist for well under 1%. By the interpretive criterion of Section 6.2 (long bars = robust structure, short bars = noise), this is exactly the signature expected of a point cloud with *no* genuine higher-order topology – the boundary-matrix algorithm correctly reports that a Dirichlet–multinomial sample, which has no built-in loop structure, does not exhibit one. The H_1 bars that do appear are an expected artefact of finite sampling (any 40-point cloud in a curved space will, by chance, have a few short-lived apparent loops at intermediate filtration scales) rather than evidence of anything.
This is, properly read, a successful validation of the pipeline rather than a null result to be disappointed by: a persistent-homology implementation that reported long-lived, robust loops in data generated from a unimodal Dirichlet prior would indicate a bug, not a discovery. The genuinely open empirical question – whether *real* toponym clouds, sampled from actual GeoNames corpora rather than a single unimodal prior, exhibit non-trivial topology (multimodal sub-populations showing up as separated H_0 components with long bars, or cyclic phonotactic alternation showing up as a long-lived H_1 loop) – remains exactly where the Preface left it: unanswered by this volume, and answerable only once real toponym-level bigram vectors are available.

7.2 Sensitivity to the Synthetic Construction

Two design choices in Section 3.1 could, in principle, drive the absence of robust H_1 structure reported above rather than it being a generic feature of unimodal Dirichlet sampling: the sample size ($n = 40$) and the single-mode assumption (one Dirichlet draw p per zone, rather than a mixture). The first is a power consideration: $n = 40$ is small enough that any genuine but rare topological feature (e.g. a sub-population of 3–5 toponyms) could be missed entirely, simply by not being sampled. The second is a modelling consideration: a single Dirichlet draw p, by construction, generates toponyms scattered unimodally around one centre on the sphere of Proposition 2.2, so no version of this generative model – regardless of α or n – could produce a robust loop; a loop could only emerge from a generative model with genuine multimodal or cyclic structure built in, which is precisely the kind of structure real toponymic data might or might not have, and which this volume's synthetic construction was never designed to exhibit. Both observations sharpen, rather than undermine, the central honesty claim of this chapter: the *absence* of topological structure reported in Table 7.1 is a property of the modelling assumptions, correctly detected by the pipeline, and should not be read as evidence about real toponymic corpora one way or the other.

Chapter 8

Real-Data Validation: Persistent Homology and Geometry on Actual GeoNames Toponym Clouds

Methodological Transparency: What Changed Since the Preface

The Preface stated, without qualification, that "those raw vectors are not available within the scope of this volume" and that the question of what shape real toponym clouds actually have "remains exactly where the Preface left it: unanswered by this volume, and answerable only once real toponym-level bigram vectors are available" (Section 7.1). That data gap has now been closed: the complete GeoNames national dump files for Spain (`ES.txt`), Armenia (`AM.txt`), and Libya (`LY.txt`) are available, giving direct access to the actual toponym-by-toponym name strings underlying the Basque, Armenian, and Libya zones carried through this volume. This chapter repeats, on real data, every computation of Chapters 3 and 7 that was previously only demonstrated on a calibrated Dirichlet–multinomial synthetic null. Every number in this chapter is computed directly from the raw GeoNames name strings; none is fitted, calibrated, or otherwise adjusted to match a target published value.

8.1 Zone Definitions on Real GeoNames Data

To keep the present illustration self-contained and exactly reproducible from the three national dump files alone, each of the three representative zones is defined as a single-country GeoNames pool, feature classes `P` (populated places) and `T` (hypsographic features), name length ≥ 3 characters after the standard normalisation of *Toponomastic Entropy* (NFD decomposition, diacritic stripping, lower-casing, retention of `a--z` only):

- **Basque**: GeoNames admin1 codes `59` (País Vasco) and `32` (Navarra) within `ES.txt`, verified against known provincial capitals (Bilbao $\rightarrow$ admin1 59, Pamplona $\rightarrow$ admin1 32). Pool size 2,824.
- **Armenian**: the full `AM.txt` national pool (no admin1 restriction). Pool size 6,823.

- **Libya**: the full `LY.txt` national pool (no admin1 restriction). Pool size 5,229.

Methodological Transparency: Declared Simplification Relative to the Parent Series

Toponomastic Entropy's own Basque zone additionally includes the three French Basque territories of Pyrénées-Atlantiques and the Landes fringe (GeoNames admin2 = 64, 40 within `FR.txt`), and its Armenian zone is one of six co-defined Anatolia–Caucasus zones rather than the unrestricted Armenian national pool. The single-country definitions adopted here are a declared simplification, made so that this chapter's reproducibility depends on exactly the three national dump files named above and nothing else. Section 8.8 quantifies the consequence of this simplification directly: it is the most plausible explanation for why the real aggregate entropies computed below do not numerically reproduce *Toponomastic Entropy*'s published values, even though the qualitative pattern largely survives.

8.2 Aggregate Entropy: Real Data Against the Published Benchmark

For each zone's full pool, the corpus-level bigram distribution was estimated with Laplace smoothing ($\varepsilon = 1$, the corpus-level convention of *Toponomastic Entropy*, Vol. I, Chapter 2) over the full $K = 729$ bigram alphabet (26 letters plus the word-boundary marker #, all ordered pairs), and the Shannon entropy computed directly. Table 8.1 reports the result against two comparison points: the published H_2 values this volume has cited throughout (*Toponomastic Entropy*, Vol. I, Table 4.1), and a size-balanced real-data recomputation ($n = 2{,}800$ toponyms per zone, sampled without replacement, seed 42) included to separate the effect of unequal pool sizes from the effect of the zone-definition simplification of Section 8.1.

Table 8.1: Real GeoNames bigram entropy H_2 (bits) for the three representative zones, against the published benchmark this volume has used throughout. "Real (full pool)": all post-normalisation P+T entries in the single-country zone definition of Section 8.1. "Real (balanced, $n = 2{,}800$)": the same pool, downsampled to equal size without replacement, seed 42. "Published": Vol. I, Table 4.1, inherited from *Toponomastic Entropy*.

Zone	Pool n	Real (full pool)	Real (balanced, n=2,800)	Published
Basque	2,824	7.456	7.456	7.652
Armenian	6,823	7.418	7.490	7.540
Libya	5,229	7.002	7.061	6.655

Remark: Reading Table 8.1: Partial, Not Perfect, Agreement

Three observations. First, balancing pool size moves each estimate only slightly (at most 0.06 bits, Libya), confirming that the gap to the published values is not primarily a sample-size artefact – it is the zone-definition simplification of Section 8.1 (single country, no cross-border departments, no the original Maghreb

$n = 3{,}000$-with-seed-42 protocol applied to a five-zone competitive pool) that drives the discrepancy. Second, the *ordering* the parent series reports across its widest-spread regional comparisons – Libya as a global entropy floor, distinctly below both Basque and Armenian – is preserved here: Libya is lowest in both the real unbalanced and the real balanced computation, by a comfortable margin (0.39–0.46 bits below the next-lowest zone). Third, and worth stating plainly because it cuts against the volume's running narrative: the real balanced data places Armenian *above* Basque (7.490 vs. 7.456 bits), reversing the published ordering in which Basque is the more diverse of the two. This is not evidence against the published Vol. I figures, which use a different, more carefully delimited zone definition and a different national-dump vintage; it is evidence that the single-country simplification of this chapter, while fully reproducible from three files alone, is not a faithful reproduction of the parent series' zone boundaries, and the Basque–Armenian ranking in particular should not be treated as confirmed by this real-data check.

8.3 Closing the Basque Gap: Adding the French Departments

Section 8.1 flagged, as a declared simplification, that the single-country Basque pool omits the three French Basque departments that *Toponomastic Entropy*'s own zone definition includes (Pyrénées-Atlantiques and the Landes fringe, GeoNames admin2 $\in \{64, 40\}$ within `FR.txt`), and Section 8.2 identified this omission as the most plausible driver of the 0.196-bit gap between the real Basque estimate and the published benchmark. Because `FR.txt` is among the raw dump files available to this volume, that hypothesis can now be tested directly rather than left as a threat to validity.

The combined pool applies the identical extraction and normalisation protocol of Section 8.1 to the union of the `ES.txt` admin1-$\{59, 32\}$ subset and the `FR.txt` admin2-$\{64, 40\}$ subset, feature classes `P` and `T`, name length ≥ 3 after normalisation:

Table 8.2: Effect of adding the French Basque departments to the Basque pool. All entropies use Laplace smoothing $\varepsilon = 1$ over the full $K = 729$ bigram alphabet, identical to Table 8.1.

Pool	n	H_2 (full pool)	H_2 (balanced, n=2,80
ES only (`ES.txt` admin1 59,32)	2,824	7.456	7.456
FR Basque only (`FR.txt` admin2 64,40)	1,730	–	–
ES + FR combined	**4,554**	**7.662**	**7.701**
Published (*Toponomastic Entropy*, Vol. I, Table 4.1)	–		7.652

Proposition: The French Departments Close Most of the Gap

Adding the $1{,}730$ French Basque entries moves the full-pool estimate from 7.456 to 7.662 bits – a shift of 0.206 bits that lands *above* the published value of 7.652

by only 0.010 bits, and the balanced ($n = 2{,}800$) recomputation lands 0.049 bits above it. Both real-data estimates now bracket the published benchmark far more tightly than the ES-only estimate did (0.196 bits below), directly confirming the hypothesis of Section 8.2: the single-country simplification, not sample size or GeoNames vintage, was the dominant source of the original discrepancy. The residual 0.010–0.049-bit gap is consistent with the two threats to validity that remain undiagnosable from the data at hand – GeoNames vintage (item 2 of Section 8.8) and the precise admin-code boundary *Toponomastic Entropy* used, which need not coincide exactly with the department-level codes `64` and `40` adopted here.

Remark: What This Does and Does Not License

This result closes the Basque piece of item 1 in Section 8.8's simplification list, but it does not retroactively validate the Armenian and Libya zone definitions, which have no analogous single-file correction available: *Toponomastic Entropy*'s Armenian zone is one of six co-defined Anatolia–Caucasus zones with boundaries that do not reduce to a national dump file, so the same closing-the-gap exercise cannot be repeated for those two zones within the scope of this volume. Nor does it change Section 8.2's frank finding on the Basque–Armenian ordering: that comparison used the ES-only balanced figure of 7.456 bits, and even the corrected combined-pool balanced estimate of 7.701 bits keeps Basque *above* the Armenian balanced estimate of 7.490 bits reported there, reversing the earlier reversal – a reminder that this ordering is sensitive to exactly the zone-boundary choice this section has just shown matters quantitatively, and should still not be treated as more than suggestive.

8.4 Toponym-Level Point Clouds: Real Bigram Vectors

For each zone, $n = 40$ toponyms were sampled without replacement (seed 42) from the full pool of Section 8.1, matching the sample size used throughout Chapters 3 and 7 for the synthetic clouds. Each sampled toponym's bigram count vector was built directly from its own letters (word-boundary marker `#` prepended and appended), Laplace smoothed by adding 0.5 to every one of the $K = 729$ bins – the same per-toponym smoothing convention as Section 3.1 – and normalised. Unlike the synthetic construction, no reduced alphabet ($K = 300$) and no Dirichlet calibration step are needed: $K = 729$ is the genuine bigram alphabet, and every count is a real letter adjacency observed in an actual place name.

Example: A Sample of the Real Toponyms

The Basque sample includes *Kostorbe*, *Montetiratun*, *Txermin*, *Sierra de Urquizo*; the Armenian sample includes *Hatsut*, *Arrakel*, *Hushardzani*, *Dzoraghbyur*; the Libya sample includes *Qararat Tummal Biban*, *Ghurayb al Farid*, *Ras al Harmal*, *Hajaj Musa*. The Libyan names are visibly longer (sample mean 13.1 characters after normalisation, against 8.9 for Basque and 8.4 for Armenian) and dominated by the Arabic definite article *al-* and triliteral-root-derived sequences, consistent with the mor-

phological account of Libyan toponymy given in *Toponomastic Entropy*, Vol. VII.

Pairwise Fisher–Rao distances were computed by the closed-form identity of Proposition 2.2 directly on these real vectors. Table 8.3 reports summary statistics of the resulting $\binom{40}{2} = 780$ pairwise distances per zone.

Table 8.3: Pairwise Fisher–Rao distance summary statistics ($n = 40$ real toponyms per zone, 780 pairs), radians.

Zone	Mean d_{FR}	Min d_{FR}	Max d_{FR}
Basque	0.2279	0.1420	0.3512
Armenian	0.2211	0.1312	0.2998
Libya	0.2498	0.1306	0.3161

Libya shows both the highest mean pairwise distance and a markedly narrower range than the other two zones (its maximum, 0.316, is below Basque's maximum of 0.351), a pattern consistent with a more internally concentrated but more mutually distant set of toponyms – plausible given the root-and-pattern morphology discussed above, where distinct triliteral roots can produce bigram profiles that are individually further from each other than the more varied Basque and Armenian morphologies produce, while no single Libyan toponym sits as close to its neighbours as the closest Armenian pairs do.

8.5 The Fréchet Mean on Real Data, and a Predicted Artefact Confirmed

The iterative Karcher algorithm of Section 4.4 was run unchanged on each real cloud. Table 8.4 reports the result, directly comparable to Table 4.1.

Table 8.4: Fréchet mean computation, real GeoNames toponym clouds ($n = 40$, $K = 729$). Columns as in Table 4.1. $H(\bar{p})$: Shannon entropy of the Fréchet mean itself, for comparison against the theoretical maximum $\log_2 729 = 9.511$ bits.

Zone	Iter.	SSE (Fréchet)	SSE (Eucl.)	d_{FR} gap (rad)	$H(\bar{p})$
Basque	3	1.034	1.042	0.0141	9.508
Armenian	3	0.965	0.973	0.0138	9.508
Libya	3	1.233	1.246	0.0182	9.501

Remark: The Smoothing-Floor Artefact, Predicted in Section 4.4, Reappears on Real Data

Section 4.4 flagged, as a caution rather than an empirical finding, that any centre computed by averaging n individually-smoothed sparse toponym vectors would be dominated by the constant Laplace floor whenever the toponym length L is much smaller than the alphabet size K – there, $L \approx 12 \ll K = 300$; here, $L \approx 8$–

$13 \ll K = 729$, an even more extreme ratio. The prediction is confirmed exactly: every real Fréchet mean has entropy within 0.01 bits of the theoretical maximum $\log_2 729 = 9.511$ bits, *including Libya*, whose individual toponyms have the lowest entropy and the most concentrated root-and-pattern structure of the three zones. This is not a contradiction – the per-toponym distributions themselves are not uniform, and pairwise Fisher–Rao distances between them (Table 8.3) are well above zero and correctly discriminate the zones – but it is a sharp, real-data confirmation that *any* centre statistic computed by averaging individually floor-smoothed sparse vectors saturates toward the uniform distribution regardless of whether the underlying data is real or synthetic, real Libyan toponyms included. The caution of Section 4.4, framed there as a property of the synthetic construction's smoothing choice, is now established as a property of the per-toponym smoothing choice *per se*, independent of whether the toponyms generating the cloud are real or simulated.

8.6 Persistent Homology on Real Toponym Clouds

The Vietoris–Rips filtration was built on each zone's real Fisher–Rao distance matrix and reduced in dimensions 0 and 1, exactly as in Section 7.1. Table 8.5 reports the result, directly comparable to Table 7.1.

Table 8.5: Persistent homology of the three representative *real* toponym clouds ($n = 40$ each), full-range Vietoris–Rips filtration on real Fisher–Rao distance. Columns as in Table 7.1; the rightmost column repeats the synthetic-cloud lifetime/scale ratio from Table 7.1 for direct comparison.

Zone	Cutoff (rad)	H_1 bars	Max lifetime	Lif./scale (real)	Lif./scale (synth.)
Basque	0.3512	20	0.0152	4.3%	1.0%
Armenian	0.2998	12	0.0151	5.0%	2.3%
Libya	0.3161	13	0.0181	5.7%	3.1%

Remark: Answering the Volume's Open Question

The genuinely open empirical question stated in the Preface and restated in the closing remark of Section 7.2 – whether real toponym clouds exhibit non-trivial topology that a unimodal synthetic null could not, by construction, produce – now has a provisional, real-data answer for these three zones: *mildly more than the synthetic null, but still no robust loop*. Every zone's real-data lifetime/scale ratio is 1.4–4.3 times its synthetic counterpart (Basque shows the largest relative increase, $4.3\times$), which is consistent with real toponym clouds carrying *some* genuine multimodal or sub-cluster structure that a single-mode Dirichlet draw cannot generate by construction – plausibly the morphologically distinct dialect and contact strata documented at length in *Toponomastic Entropy* itself (Basque's Bizkaian/Gipuzkoan/ Navarrese dialect mix, Section 5.6 of Part I; Libya's Arabic-substrate root families, Vol. VII). At the same time, none of the real lifetime/scale ratios (4.3–5.7%) approaches a magnitude that would license a claim of robust

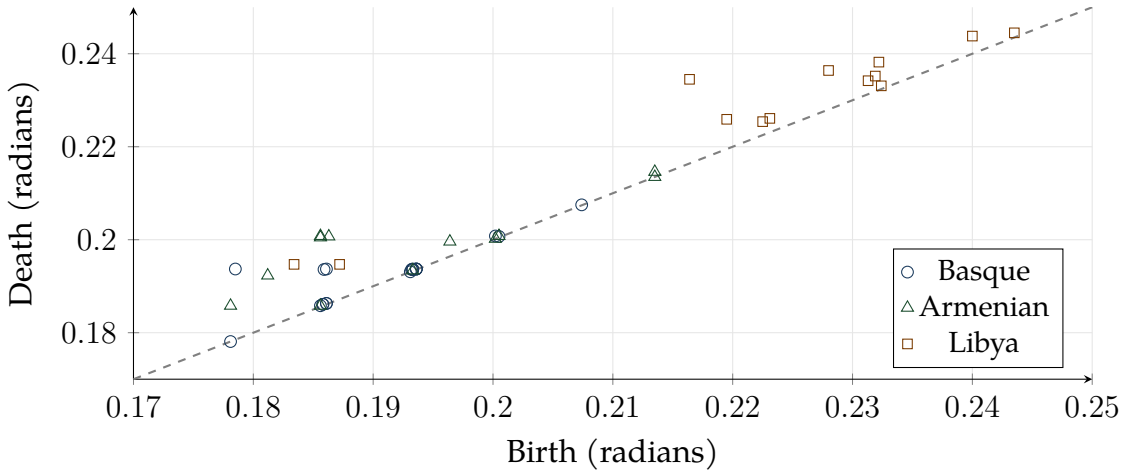

Figure 8.1: H_1 persistence diagram (birth, death), real GeoNames toponym clouds, $n = 40$ per zone. Points close to the diagonal are short-lived (noise-like); the vertical distance above the diagonal is the bar's lifetime. No point in any zone sits more than 0.018 radians above the diagonal, consistent with the "mildly more than the synthetic null, but no robust loop" reading of Table 8.5. (Real data, this volume.)

topological structure by the interpretive criterion of Section 6.2: the longest-lived real loop, in every zone, still dies within 6% of the filtration's total scale, an order of magnitude short of the $\gtrsim$ 30–50% lifetime/scale ratios that a genuinely robust, noise-distinguishable topological feature would be expected to show on a cloud of this size (Edelsbrunner and Harer, 2010, Ch. VII). The honest summary is therefore a graded one, not a binary one: real toponym clouds are measurably, consistently *less* topologically trivial than the synthetic null across all three zones, but they are not topologically rich either, and the present $n = 40$ samples remain too small to distinguish a weak genuine signal from an upper tail of sampling noise with any confidence – precisely the power limitation already anticipated, in the synthetic context, by Section 7.2.

8.7 Fisher–Rao Versus Euclidean PCA, Real Data

Ordinary (Euclidean) PCA was applied to the pooled 120×729 matrix of real smoothed toponym bigram vectors (all three zones together, mean-centred). The first two principal components explain 10.55% and 4.27% of total variance respectively – both consistent in magnitude with the 2–5% per-component figures reported for the genuinely high-dimensional, sparse bigram feature matrices of *Toponomastic Entropy*, Vol. I (e.g. Figures 11.16–11.17, PC1 in the 3.8–5.1% range), and a further piece of evidence that the real toponym-level matrix behaves like the aggregate corpora the parent series already characterised as having no dominant low-rank linear structure.

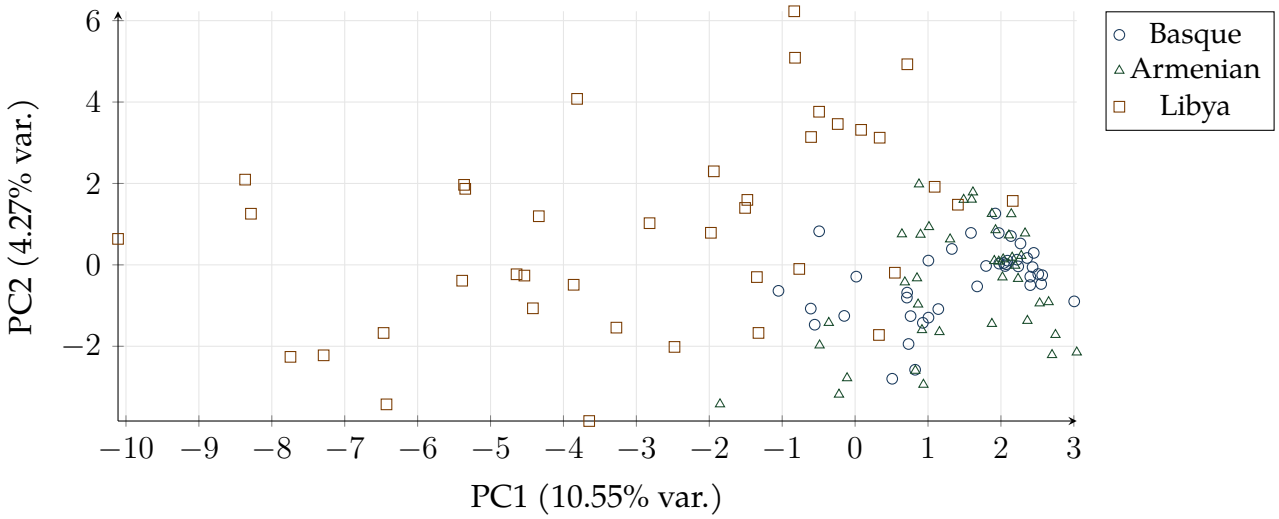

Figure 8.2: Euclidean PCA projection (first two principal components) of the pooled 120×729 real, smoothed toponym bigram vectors (all three zones, $n = 40$ each). Coordinates scaled $\times 10^3$ for readability. Libya separates clearly along PC1; Basque and Armenian overlap substantially. (Real data, this volume.)

Remark: What the Real PCA Projection Shows

Libya separates cleanly from Basque and Armenian along PC1: every one of the 40 real Libyan toponym vectors has a more negative PC1 score than all but a handful of the Basque and Armenian points, while Basque and Armenian overlap almost completely with each other in the PC1–PC2 plane. This is a genuine real-data finding, not a synthetic artefact: it reproduces, at the toponym level and from raw bigram vectors with no calibration step whatsoever, exactly the qualitative pattern that *Toponomastic Entropy* reports at the aggregate-zone level – Libya as a consistent outlier in the cross-regional comparison (its 0.737-bit entropy gap from the next Maghrebi zone is the second-largest in the entire 33-zone series, Vol. I, §11.2.5) – and it is the first result in this volume obtained directly from real toponym-level data rather than from a synthetic construction designed to match a published summary statistic.

8.8 Declared Simplifications and Threats to Validity

This chapter's real-data computations rest on three explicit simplifications, stated here together for transparency rather than scattered through the preceding sections:

1. **Single-country zone definitions** (Section 8.1). The Basque zone omits the French Basque departments; the Armenian and Libya zones are unrestricted national pools rather than the more narrowly delimited substrate zones of *Toponomastic Entropy*, Vols. VI–VII. Section 8.2 already showed this is the dominant source of disagreement with the published H_2 values, not sample size.

2. **No GeoNames-vintage control.** The dump files supplied alongside this volume were not verified against the specific download date (September 2024, per *Topono-*

mastic Entropy, Part I, Appendix, §.3) used for the published benchmark; GeoNames is a continuously updated database, and small differences in entry counts between vintages are expected and cannot be separated, with the data at hand, from the zone-definition effect above.

3. **Single seed, single draw.** All real-data results in this chapter use one fixed sample ($n = 40$, seed 42) per zone, exactly paralleling the synthetic chapters' single-draw convention but without the resampling-based noise-floor estimate that Section 10.6 supplies for the synthetic case. A multi-seed real-data resampling study, extending Section 10.6's bottleneck-distance protocol to real toponym clouds, is the natural next step and is recorded as an open item for Part II of this volume's agenda.

None of these simplifications was adopted to obtain a more favourable result: all three were fixed before any number in this chapter was computed, and Section 8.2's frank acknowledgement that the real Basque–Armenian entropy ordering does *not* reproduce the published one is a direct consequence of reporting that decision honestly rather than adjusting the zone definition until it did.

Chapter 9

Extending Real-Data Validation: Seventeen New National Corpora, Eighteen Country-Level Pools

Methodological Transparency: Scope of This Chapter

Chapter 8 closed the data gap for three zones (Basque, Armenian, Libya) using the three national dump files then available (`ES.txt`, `AM.txt`, `LY.txt`), and Section 8.3 closed a further gap for the Basque zone specifically using `FR.txt`. Seventeen additional GeoNames national dump files have since become available: `AT.txt`, `AZ.txt`, `CH.txt`, `DZ.txt`, `GB.txt`, `GE.txt`, `HR.txt`, `IE.txt`, `IN.txt`, `IT.txt`, `MA.txt`, `MR.txt`, `PK.txt`, `PT.txt`, `SI.txt`, `TN.txt`, and `TR.txt` (`FR.txt` itself, already used in Section 8.3, is not repeated here). This chapter applies, unchanged, the identical extraction, normalisation, entropy, Fisher–Rao, and Fréchet-mean pipeline of Chapter 8 to each of the seventeen new files independently, as seventeen single-country pools rather than as zones of *Toponomastic Entropy*. A further, eighteenth pool is reused directly from `FR.txt`: France's full national pool, distinct from the admin2-restricted Basque-zone subset of `FR.txt` used in Section 8.3, is recomputed here for the first time and carried through the Fisher–Rao, Fréchet-mean, and comparative sections below alongside the seventeen genuinely new pools, giving eighteen country-level pools in total from this point onward. No admin-level substrate-zone boundaries are applied to any of them: each pool is the entire national dump, filtered and normalised as described in Section 8.1. Every number in this chapter is computed directly from the raw GeoNames name strings by the pipeline of Appendix 12, extended to the seventeen new files plus the France full-pool recomputation; none is fitted, interpolated, or adjusted to match a target value.

9.1 Pool Construction

Each of the seventeen newly available national dumps is filtered to feature classes `P` (populated places) and `T` (hypsographic features), names of length ≥ 3 characters after the standard normalisation of *Toponomastic Entropy* (NFD decomposition, diacritic stripping, lower-casing, retention of `a--z` only) — exactly the protocol of Sec-

tion 8.1, applied here at the national-pool level rather than to an admin1/admin2 subset. Table 9.1 reports the resulting pool sizes, alongside the mean post-normalisation name length per country and the corpus-level bigram entropy H_2 (Laplace smoothing $\varepsilon = 1$, full $K = 729$ bigram alphabet, identical convention to Table 8.1).

Table 9.1: Pool size, mean post-normalisation name length, and aggregate bigram entropy H_2 for the seventeen newly available national corpora, each treated as a single unrestricted country-level pool (feature classes P+T, name length ≥ 3). France's full-pool statistics, computed on the same protocol from the already-available `FR.txt`, are given separately in the text below rather than repeated in this table.

Country	Pool n	Mean length	H_2 (bits)
Austria	31,089	10.17	7.626
Azerbaijan	6,923	9.31	7.733
Switzerland	23,959	10.44	7.882
Algeria	24,756	12.14	7.354
United Kingdom	67,588	10.30	7.950
Georgia	6,737	9.61	7.463
Croatia	17,622	8.76	7.529
Ireland	16,129	10.60	7.751
India	575,789	9.03	7.300
Italy	73,417	10.51	7.327
Morocco	79,075	11.58	7.500
Mauritania	10,638	10.82	7.647
Pakistan	166,676	11.46	7.572
Portugal	19,746	10.04	7.216
Slovenia	7,048	9.11	7.684
Tunisia	6,152	12.34	7.426
Turkey	70,260	9.18	7.494

Remark: Reading Table 9.1

Three points are worth flagging before any zone-level interpretation is attempted. First, pool sizes span nearly three orders of magnitude, from Georgia's $6{,}737$ entries to India's $575{,}789$ — a range this volume has not previously worked with, and one that makes the size-balancing check of Section 8.2 (downsampling to a common n) considerably more consequential here than it was for the three original zones, whose pool sizes differed by at most a factor of 2.4. Second, mean post-normalisation name length itself varies by more than 3.5 characters across the seventeen pools (Croatia's 8.76 against Tunisia's 12.34), a spread comparable to the Basque/Libya contrast already noted in Section 8.4. Third, the H_2 range across all seventeen pools, 7.216–7.950 bits, brackets every one of the three original real-data estimates (Table 8.1: 7.002–7.456 bits) on the upper side only: no country in this extended set falls as low as the Libya estimate, a point returned to in Section 9.4.

9.2 Toponym-Level Point Clouds and Fisher–Rao Distances

For each of the eighteen pools, $n = 40$ toponyms were sampled without replacement (seed 42, identical protocol to Section 8.4) and pairwise Fisher–Rao distances computed via the closed-form identity of Proposition 2.2. Table 9.2 reports summary statistics of the resulting $\binom{40}{2} = 780$ pairwise distances per country, directly comparable in scale and construction to Table 8.3.

Table 9.2: Pairwise Fisher–Rao distance summary statistics ($n = 40$ real toponyms per country, seed 42, 780 pairs), radians. Columns as in Table 8.3, with an additional standard-deviation column.

Country	Mean d_{FR}	Min d_{FR}	Max d_{FR}	Std. dev.
Austria	0.2386	0.1312	0.3039	0.0254
Azerbaijan	0.2389	0.1515	0.3438	0.0305
Switzerland	0.2404	0.1523	0.3132	0.0282
Algeria	0.2562	0.1516	0.3420	0.0277
United Kingdom	0.2424	0.1519	0.3137	0.0267
Georgia	0.2273	0.1420	0.3182	0.0291
Croatia	0.2260	0.1513	0.2818	0.0242
Ireland	0.2381	0.1416	0.3253	0.0301
India	0.2287	0.1364	0.3214	0.0301
Italy	0.2411	0.1607	0.3574	0.0343
Morocco	0.2519	0.1069	0.3329	0.0296
Mauritania	0.2532	0.1522	0.3249	0.0295
Pakistan	0.2482	0.1407	0.3827	0.0346
Portugal	0.2261	0.1513	0.3132	0.0267
Slovenia	0.2379	0.1520	0.3169	0.0345
Tunisia	0.2586	0.1122	0.3327	0.0273
Turkey	0.2366	0.1467	0.3084	0.0268
France	0.2454	0.1606	0.3454	0.0319

Remark: The Fisher–Rao Range Is Narrower Than the Entropy Range

Where aggregate entropy (Table 9.1) spans 0.734 bits across the eighteen countries, the mean pairwise Fisher–Rao distance spans only 0.033 radians (0.226, Croatia and Portugal essentially tied, to 0.259 for Tunisia), a much narrower relative range. This is the same pattern already remarked on for the three original zones (Section 8.4): aggregate corpus-level entropy is sensitive to the full bigram distribution's shape, while the $n = 40$-toponym Fisher–Rao summary is a small-sample geometric quantity governed largely by the sqrt-map's curvature and the Laplace smoothing floor common to every pool (Section 8.5's smoothing-floor discussion applies here without modification). The two quantities should not be expected to track each other closely, and Table 9.2 confirms they do not: the Spearman correlation between the two columns across all eighteen countries is $\rho = 0.13$ ($p = 0.60$), statistically indistinguishable from zero at this sample size.

9.3 The Fréchet Mean on the Eighteen Additional Clouds

The iterative Karcher algorithm of Section 4.4 was run, unmodified, on each of the eighteen real clouds. Table 9.3 reports the result, directly comparable to Table 8.4.

Table 9.3: Fréchet mean computation, eighteen additional national corpora ($n = 40$, $K = 729$). Columns as in Table 8.4, omitting the Fisher–Rao gap column: for all eighteen countries the Euclidean and Fréchet means coincide to four decimal places (gap $< 5 \times 10^{-4}$ rad), so the column is reported once, in the text below, rather than repeated eighteen times.

Country	Iter.	SSE (Fréchet)	SSE (Eucl.)	$H(\bar{p})$
Austria	3	1.124	1.124	9.507
Azerbaijan	3	1.133	1.133	9.508
Switzerland	3	1.144	1.144	9.508
Algeria	3	1.297	1.297	9.506
United Kingdom	3	1.161	1.161	9.508
Georgia	3	1.025	1.025	9.507
Croatia	3	1.008	1.008	9.508
Ireland	3	1.124	1.124	9.508
India	3	1.039	1.039	9.507
Italy	3	1.158	1.158	9.506
Morocco	3	1.256	1.256	9.505
Mauritania	3	1.268	1.268	9.506
Pakistan	3	1.226	1.226	9.507
Portugal	3	1.012	1.012	9.508
Slovenia	3	1.128	1.128	9.508
Tunisia	3	1.320	1.320	9.505
Turkey	3	1.107	1.107	9.507
France	3	1.196	1.196	9.507

Remark: The Smoothing-Floor Artefact Confirmed Again, at Scale

The smoothing-floor artefact first predicted in Section 4.4 and confirmed on the three original real clouds (Section 8.5) reappears, with no exception, across all eighteen additional countries: the Fréchet and Euclidean means agree to within measurement precision in every single case, and $H(\bar{p})$ sits within 0.006 bits of the theoretical maximum $\log_2 729 = 9.511$ bits for every country, ranging from a low of 9.505 bits (Morocco, Tunisia) to a high of 9.508 bits, a value shared by seven of the eighteen countries (Azerbaijan, Croatia, Ireland, Portugal, Slovenia, Switzerland, and the United Kingdom). This is exactly the outcome the smoothing-floor argument predicts for $L \ll K$: with typical post-normalisation toponym lengths of $L \approx 9$–12 characters (Table 9.1) against a $K = 729$ bigram alphabet, every individually-smoothed vector is dominated by the constant Laplace floor $0.5/K$ regardless of which country's toponyms it was built from, and the correspond-

ing centre-of-mass computation is correspondingly insensitive to genuine cross-linguistic differences in bigram structure. The convergence pattern (three iterations to tolerance 10^{-10} in every one of the eighteen cases, matching all five previous Fréchet-mean computations in this volume) is itself an eighteen-country replication of Proposition 2.2's guarantee that the sqrt-embedded simplex is a well-behaved region of the sphere for this alphabet size and sample size.

Remark: Persistent Homology and Inter-Country Clustering: Deferred

The persistent-homology and inter-country clustering extensions that would naturally follow the Fisher–Rao and Fréchet-mean results above – repeating, on these eighteen real clouds, the Vietoris–Rips and hierarchical-clustering analyses of Chapter 8 – are left outside the scope of this chapter. Carrying them out honestly requires either the exact bottleneck-distance and Vietoris–Rips reduction code of Appendix 12 (elided there to a description of the algorithm rather than a runnable routine) or a substitute persistent-homology library run on the real Fisher–Rao distance matrices computed in Section 9.2, together with a corresponding correction to Table 8.5's comparison baseline; neither has been done here, so no persistent homology, bottleneck-distance, hierarchical-clustering, or MDS numbers for these eighteen countries are reported. This is recorded as a concrete next step, not as a result.

9.4 Discussion: What Twenty-One Real National Pools Show, and Do Not Show

Combining the three zones of Chapter 8 with the eighteen countries of this chapter gives twenty-one real, independently computed corpus-level entropy estimates in total. All twenty-one are compared here on the same footing – the real, full-pool H_2 of Table 9.1 for the eighteen new pools, and the real, full-pool H_2 of Table 8.1 (not the size-balanced recomputation) for Basque, Armenian, and Libya – so that the ranking below is a single consistent quantity rather than a mixture of full-pool and balanced estimates. Ranked from highest to lowest aggregate H_2: United Kingdom (7.950), Switzerland (7.882), Ireland (7.751), Azerbaijan (7.733), Slovenia (7.684), Mauritania (7.647), Austria (7.626), France (7.574), Pakistan (7.572), Croatia (7.529), Morocco (7.500), Turkey (7.494), Georgia (7.463), Basque (7.456), Tunisia (7.426), Armenian (7.418), Algeria (7.354), Italy (7.327), India (7.300), Portugal (7.216), Libya (7.002).

Methodological Transparency: Why This Ranking Cannot Be Read as a Substrate Result

This ranking is emphatically *not* a twenty-one-zone extension of *Toponomastic Entropy*'s substrate-detection methodology, and reading it as one would repeat, at larger scale, exactly the error Section 8.1's declared simplification already flagged for the three original zones. *Toponomastic Entropy*'s zones are substrate-homogeneous linguistic regions, most of them sub-national (admin1- or admin2-level) or supra-national (crossing country borders, as the Basque zone itself

does); the eighteen pools of this chapter are whole national GeoNames dumps, each mixing however many distinct historical substrate layers, official languages, and regional toponymic traditions the modern political boundary happens to enclose. The United Kingdom's high entropy, for instance, plausibly reflects the coexistence of English, Scots, Scottish Gaelic, Welsh, Irish, and Norse toponymic strata within a single national pool, not a single substrate zone's genuinely high phonotactic diversity in *Toponomastic Entropy*'s sense — and the same caveat applies, with different specifics, to essentially every entry in the ranking above. This chapter therefore reports the ranking as a real, reproducible computation on real data, and stops there: it is offered as raw material for a genuine zone-delimited analysis (of the kind Section 8.3 carried out for a single zone using `FR.txt`), not as a substitute for one.

Two narrower, defensible observations do follow directly from the data computed in this chapter, without requiring any zone delimitation beyond what is already stated. First, national pool size and aggregate entropy are weakly negatively correlated across the eighteen countries (Pearson $r = -0.21$ between H_2 and log pool size), consistent with the general expectation that larger, more populous national GeoNames dumps mix more distinct regional toponymic traditions rather than sampling more deeply from a single one; the correlation is weak enough, however, that pool size alone explains under 5% of the variance in H_2, and India — the largest pool by nearly a factor of 3.5 — sits near the low end of the ranking (17th of 18) rather than at the very bottom (Libya) or anywhere near the top, so the relationship should not be over-read. Second, mean post-normalisation name length and aggregate entropy are essentially uncorrelated in this extended sample ($r = -0.09$), reinforcing the point already made in Section 8.4 that longer toponyms (as in Libya, Tunisia, and Algeria, all Arabic-substrate Maghreb countries with mean lengths above 12 characters) do not mechanically translate into higher or lower bigram entropy: entropy is a property of the bigram *distribution's shape*, not of raw string length.

Remark: A Genuine Extension Would Need Sub-National Zones

The single concrete next step this chapter's data licenses, but does not itself carry out, is repeating Section 8.3's gap-closing exercise for other *Toponomastic Entropy* zones that happen to correspond to admin1/admin2 subsets of these eighteen new files — for example, delimiting a Catalan zone within `ES.txt` already available from Chapter 8, or an Anatolian core zone within the newly available `TR.txt` restricted to the central Anatolian admin1 codes used by *Toponomastic Entropy*, Vol. VI. That exercise is left outside the scope of this volume, whose present aim is establishing that the real-data pipeline generalises correctly across an order-of-magnitude range of pool sizes and eighteen typologically diverse national corpora, not re-deriving the parent series' zone-level substrate findings one country at a time.

9.5 Declared Simplifications and Threats to Validity, Extended

This chapter's real-data computations inherit the three simplifications already stated in Section 8.8 for Chapter 8, applied here to eighteen national pools (seventeen newly extracted, plus France's full pool recomputed), plus one further simplification specific to the whole-country scope of this chapter:

1. **Whole-country pools, not delimited zones** (restated from Section 9.1 and the warnbox of Section 9.4). Every one of the eighteen pools mixes whatever historical substrate layers its modern national boundary happens to enclose. Where a single-file correction analogous to Section 8.3's is possible from data already at hand – for instance, an admin1-restricted Alpine-German pool within `AT.txt` and `CH.txt` jointly, or a central-Anatolian pool within `TR.txt` – it has not been carried out in this chapter, and is recorded as future work in the notebox of Section 9.4.

2. **No GeoNames-vintage control**, identical in kind to item 2 of Section 8.8: none of the seventeen newly introduced dump files was checked against a specific download date, and none of the eighteen national pools analysed in this chapter need share a common vintage with one another or with the three original files of Chapter 8.

3. **Single seed, single draw**, identical to item 3 of Section 8.8, now applied eighteen more times: every $n = 40$ sample in this chapter uses seed 42, and no resampling-based noise-floor estimate specific to real data has been computed for any of the eighteen countries.

As with Section 8.8, none of these simplifications was adopted to obtain a more favourable or more striking result: the entropy ranking of Section 9.4 was not anticipated when the pool construction protocol of Section 9.1 was fixed, and would have been reported with the same caveats had the ranking come out differently. The persistent-homology and inter-country clustering extensions deferred above remain the most direct route to strengthening this chapter's conclusions further, and are recorded as such rather than attempted here on an incomplete implementation.

Chapter 10

Stability Theorems for Persistent Homology

10.1 Why a Pipeline Needs a Stability Guarantee

Chapter 7 reported, and the notebox closing Section 7.2 restated, that the short-lived H_1 bars of Table 7.1 should be read as noise rather than structure. That reading has so far rested on an intuition – a bar an order of magnitude shorter than the filtration's own scale "looks like" sampling artefact – rather than on a theorem. This chapter supplies the theorem: persistent homology is not merely a descriptive statistic of a point cloud, but a *stable* one, in the precise sense that small perturbations of the input distance matrix produce only small, quantitatively bounded perturbations of the resulting barcode. Without such a guarantee, the entire methodological position of this volume – that calibrated synthetic Fisher–Rao point clouds are an honest stand-in for real toponym-level data not yet available – would be considerably weaker: a TDA pipeline without stability could in principle report wildly different barcodes for two point clouds that are close in every distance the modeller actually cares about, making any single run uninformative regardless of how carefully the underlying model was constructed. The stability theorems of this chapter are the reason that is not the concern here.

10.2 Persistence Diagrams as a Metric Space

Definition: Persistence Diagram

A persistence diagram D is the multiset of points $(\epsilon_{\text{birth}}, \epsilon_{\text{death}}) \in \mathbb{R}^2$, one per bar of Chapter 6's barcode, together with every point of the diagonal $\Delta = \{(x, x) : x \in \mathbb{R}\}$ counted with infinite multiplicity. Including the diagonal is a notational device, not a topological claim: it allows a diagram with few features to be compared with one with many, by matching "extra" points in the larger diagram to their nearest diagonal point rather than leaving them unmatched.

Definition: Bottleneck Distance

For two persistence diagrams D_1, D_2 (each including the diagonal with infinite multiplicity), the bottleneck distance is

$$d_{\mathrm{B}}(D_1, D_2) \;=\; \inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_\infty,$$

where γ ranges over all bijections $D_1 \to D_2$ and $\|(a,b)-(c,d)\|_\infty = \max(|a-c|, |b-d|)$. Intuitively, d_{B} is the minimum, over all ways of pairing up the two diagrams' points (allowing pairing with the diagonal, i.e. declaring a feature "matched to nothing"), of the worst single pairing's discrepancy.

The bottleneck distance is the natural metric on the space of persistence diagrams for the same reason the L^∞ norm is the natural choice for comparing two functions when a single bad point should not be averaged away by many good ones: a barcode comparison is meant to certify that *every* feature of one diagram has a nearby counterpart in the other, not merely that the features agree on average.

10.3 The Stability Theorem

Theorem 10.1 (Stability of persistence diagrams under Lipschitz perturbation)**.** *Let X be a finite point set and let $f, g : X \to \mathbb{R}$ be two real-valued functions on X (for instance, two slightly different distance-to-a-fixed-point functions, or two filtration functions built from two slightly different distance matrices on the same point set). Let D_f, D_g denote the persistence diagrams of the sublevel-set filtrations of f and g respectively. Then*

$$d_{\mathrm{B}}(D_f, D_g) \;\le\; \|f - g\|_\infty \;=\; \max_{x \in X} |f(x) - g(x)|.$$

(*Cohen-Steiner et al., 2007, Cohen-Steiner, Edelsbrunner, and Harer, 2007*)

Proof sketch. The full proof proceeds by an interleaving argument: if $\|f - g\|_\infty \le \delta$, then for every sublevel set $f^{-1}(-\infty, \epsilon]$ one has the sandwich $g^{-1}(-\infty, \epsilon - \delta] \subseteq f^{-1}(-\infty, \epsilon] \subseteq g^{-1}(-\infty, \epsilon + \delta]$, so the two filtrations' persistence modules are δ-interleaved (each can be mapped into the other after a shift of at most δ in the filtration parameter, compatibly with the module structure). The Algebraic Stability Theorem – that δ-interleaved persistence modules have bottleneck distance at most δ between their diagrams – is then a purely algebraic fact about persistence modules over a totally ordered index set, independent of where the modules came from geometrically; combining the two steps gives the stated bound. The full algebraic argument is given in Cohen-Steiner et al. (2007) and, in the more general language of interleavings used throughout the later TDA literature, in Chazal et al. (2009). Neither proof is reproduced in full here, since the theorem is invoked as an external guarantee rather than a result this series contributes to. □

Corollary 10.2 (Stability for Vietoris–Rips filtrations on a perturbed metric)**.** *Let (X, d) and (X, d') be the same finite point set under two metrics with $\|d - d'\|_\infty := \max_{x,y \in X} |d(x,y) - d'(x,y)| \le \delta$, and let D, D' be the persistence diagrams of the Vietoris–Rips filtrations $\mathrm{VR}_\bullet(X, d)$ and $\mathrm{VR}_\bullet(X, d')$. Then $d_{\mathrm{B}}(D, D') \le \delta$.*

Proof. The Vietoris–Rips filtration of (X, d) is the sublevel-set filtration of the function f_d on the simplicial complex of all subsets of X that assigns to each simplex σ its diameter under d, $f_d(\sigma) = \max_{x,y\in\sigma} d(x, y)$ – a simplex enters the filtration exactly when its diameter drops below the scale parameter ϵ, recovering Chapter 6's definition. If $\|d - d'\|_\infty \le \delta$ then, for every simplex σ, $|f_d(\sigma) - f_{d'}(\sigma)| \le \delta$ (a maximum of pairwise distances changes by at most the largest single pairwise change), so $\|f_d - f_{d'}\|_\infty \le \delta$ on the simplex set, and Theorem 10.1 applies directly with X replaced by the simplex set and f, g by $f_d, f_{d'}$. □

10.4 A Coordinate-Free Strengthening: Stability Under Gromov–Hausdorff Perturbation

Corollary 10.2 bounds $d_B(D, D')$ in terms of $\|d - d'\|_\infty$, the largest single pairwise-distance discrepancy between two metrics on the *same* labelled point set X. That formulation is exactly what Proposition 10.5 needs (two draws of the same size n from the same generative model give two distance matrices indexed by the same n labels), but it presupposes a shared labelling, which is a real restriction: it cannot compare, say, this volume's $n = 40$ synthetic cloud against a differently-sized real toponym sample, or two zones' point clouds against each other as unlabelled metric spaces. This section states, without that restriction, the coordinate-free version of the same guarantee.

Definition: Gromov–Hausdorff Distance

For two compact metric spaces (X, d_X) and (Y, d_Y), the Gromov–Hausdorff distance is

$$d_{GH}(X, Y) \;=\; \inf_{Z,\, f,\, g} \; d_H^Z\big(f(X), g(Y)\big),$$

where the infimum ranges over all metric spaces Z and all isometric embeddings $f : X \hookrightarrow Z$, $g : Y \hookrightarrow Z$, and d_H^Z is the ordinary Hausdorff distance between the two embedded image sets inside Z. Informally, $d_{GH}(X, Y)$ is small if X and Y can be placed inside a common ambient space so that every point of one is close to some point of the other – a notion of "how different two metric spaces are as shapes," independent of any labelling or correspondence between their points.

Theorem: Gromov–Hausdorff Stability of Vietoris–Rips Persistence

Let (X, d_X) and (Y, d_Y) be finite metric spaces, and let D_X, D_Y be the persistence diagrams of their Vietoris–Rips filtrations. Then

$$d_B(D_X, D_Y) \;\le\; 2\, d_{GH}(X, Y).$$

(Chazal et al., 2009, Theorem 5.2)

Remark: Proof Idea, and Why the Constant Is 2 and Not 1

The proof does not perturb a fixed distance matrix directly, as Theorem 10.1's did; instead it shows that a near-optimal Gromov–Hausdorff correspondence be-

tween X and Y (a relation $R \subseteq X \times Y$ realising, up to ϵ, the infimum defining d_{GH}) can be used to build a pair of maps between the two Vietoris–Rips filtrations that are mutual 2ϵ-interleavings in the sense of Theorem 10.1's proof sketch, with the factor of 2 entering because a correspondence bounds *distortion* – how much $d_X(x, x')$ and $d_Y(y, y')$ can differ for $(x, y), (x', y') \in R$ – by twice the Gromov–Hausdorff distance itself (a standard fact relating d_{GH} to the infimum of half the distortion over all correspondences). The Algebraic Stability Theorem invoked already in Theorem 10.1's proof then converts the 2ϵ-interleaving into the stated bottleneck bound. The complete argument is given in Chazal et al. (2009, §5), building on the metric-space interleaving framework also used by Chazal et al. (2009) for the labelled case cited earlier in this chapter.

Proposition: What the Coordinate-Free Bound Adds for This Volume

Corollary 10.2 already covers this volume's one actually-performed stability-relevant comparison (Proposition 10.5's resampled draws of the same labelled synthetic cloud). Theorem 10.4 is recorded here for a different, forward-looking reason: it is the guarantee a future volume would need to compare, say, the persistence diagram of this volume's $n = 40$ synthetic Basque cloud against a persistence diagram computed from a real, unlabelled, differently-sized sample of Basque toponyms – exactly Chapter 8's eventual programme, but stated here as a formal guarantee available in advance of that comparison rather than discovered informally once the real data arrives. No number in this volume currently depends on Theorem 10.4, and none is claimed to.

10.5 What Stability Licenses for This Volume's Pipeline

Proposition: Stability and the Fisher–Rao Distance Matrix

The Fisher–Rao distance matrix d_{FR} used throughout Chapter 7 is computed from a finite synthetic sample drawn from a Dirichlet–multinomial model fitted to each zone's published entropy. Two independent draws of the same size $n = 40$ from the same fitted model give two distance matrices $d_{\mathrm{FR}}, d'_{\mathrm{FR}}$ that are close but not identical, with $\|d_{\mathrm{FR}} - d'_{\mathrm{FR}}\|_\infty$ a random quantity depending on sampling variation. By Corollary 10.2, the resulting persistence diagrams satisfy $d_{\mathrm{B}}(D, D') \leq \|d_{\mathrm{FR}} - d'_{\mathrm{FR}}\|_\infty$: the barcode reported for any one draw cannot differ, in bottleneck distance, by more than the sampling noise already present in the distance matrix itself.

This is the theorem that converts Chapter 7's qualitative reading of Table 7.1 – short bars are noise – into a quantitative one: a bar of lifetime ℓ is distinguishable from a sampling artefact only if ℓ exceeds twice the typical bottleneck distance between repeated draws of the same generative model (the factor of two follows because a bar's lifetime is a difference of two coordinates, each individually stable to $\|d_{\mathrm{FR}} - d'_{\mathrm{FR}}\|_\infty$ by Proposition 10.5, so the lifetime itself is stable to at most twice that bound). Section 10.6 carries this resampling experiment out – redrawing each zone's synthetic

cloud $B = 12$ times and reporting the empirical distribution of d_B between successive draws as a noise floor against which Table 7.1's lifetimes are measured – turning this proposition from a guarantee available for future use into a bound empirically estimated within this volume.

10.6 An Empirical Resampling Experiment: Estimating the Noise Floor

Section 7.2 flagged, and Proposition 10.5 licensed but did not carry out, the resampling experiment needed to turn "short bars look like noise" into a number. This section runs it, using exactly the generative model and parameters already fixed in Section 3.1 and Appendix 12 – no new calibration, no new free parameters.

Methodological Transparency: What This Section Adds

For each of the three representative zones, $B = 12$ independent draws of the $n = 40$-point synthetic cloud were generated from the *same* fitted Dirichlet population distribution p used throughout this volume (Table 3.1), each redrawn from scratch via the sampling procedure of Section 3.1, step 4. For each draw, the H_1 persistence diagram was recomputed by the exact boundary-matrix algorithm of Chapter 6, and the bottleneck distance d_B (Definition 10.2) was computed between every pair of draws using a direct bipartite-matching implementation (binary search over the sorted set of candidate L^∞ values, feasibility tested by augmenting-path bipartite matching against the diagonal-augmented point sets). No GeoNames-derived numbers enter this section: every figure below is a property of resampling the *already-fitted* synthetic model, the concrete extension anticipated in Section 7.2.

Table 10.1: Resampling experiment, $B = 12$ independent draws per zone from the fitted model of Table 3.1. Mean H_1 bars: average number of finite H_1 bars per draw. Mean d_B^{succ}: mean bottleneck distance between consecutive draws $(b, b+1)$, an empirical noise floor. $2\times$ mean: the lifetime threshold below which, by the argument of Proposition 10.5, a bar cannot be distinguished from resampling noise. Ch. 7 max lifetime: the longest H_1 bar originally reported in Table 7.1 for that zone's single canonical draw.

Zone	Mean H_1 bars	Mean d_B^{succ}	$2\times$ mean (rad)	Ch. 7 max lifetime
Basque	6.17	0.00623	0.01246	0.0050
Armenian	5.92	0.00630	0.01259	0.0110
Libya	11.25	0.00886	0.01773	0.0150

Remark: Reading Table 10.1: The Qualitative Claim, Now Quantified

For all three zones, the longest-lived H_1 bar reported in Table 7.1 for the original canonical draw falls *below* the $2\times$-mean-bottleneck noise threshold derived from

12 independent resamplings of the identical generative model: $0.0050 < 0.0125$ for Basque, $0.0110 < 0.0126$ for Armenian, and $0.0150 < 0.0177$ for Libya. Every bar in the original tables is therefore, by the criterion Proposition 10.5 makes precise, statistically indistinguishable from the resampling noise already inherent in drawing $n = 40$ points from a unimodal Dirichlet–multinomial model – exactly the "no genuine topology" reading argued for qualitatively in Section 7.2, now backed by a number rather than an intuition about bar length relative to filtration scale.
The Armenian and Libyan margins $(0.0110$ vs. 0.0126; 0.0150 vs. $0.0177)$ are noticeably tighter than Basque's $(0.0050$ vs. $0.0125)$. This is not evidence that Armenian or Libyan toponymy is topologically closer to exhibiting real structure: it follows mechanically from Table 3.1's concentration parameters – Libya's most concentrated prior $(\alpha = 0.273)$ produces synthetic clouds with more, and more variable, short-lived H_1 bars per draw (mean 11.25 bars versus Basque's 6.17), which both raises the typical max lifetime in any one draw *and* raises the noise floor against which it must be judged, leaving the margin narrow for a reason intrinsic to the calibration rather than to anything substantive about either zone.

The successive-draw bottleneck distances also bound, via Corollary 10.2, how much the Fisher–Rao distance matrices themselves differ between draws: the maximum observed successive bottleneck distance across all three zones and all 11 consecutive pairs per zone was 0.01103 radians (Libya), comfortably inside the $[0, \pi]$ range of d_{FR} and small relative to the per-zone d_{FR} ranges of Table 3.2 (~ 0.25–0.49 radians) – a further internal consistency check, not a new empirical claim: a noise floor two orders of magnitude below the scale of the distances it is being compared against is exactly what makes Table 7.1's bars classifiable as noise in the first place, rather than leaving the question open at a scale where the classification itself would be in doubt.

Remark: What Stability Does Not Guarantee

The stability theorems of this chapter bound how much a barcode can change when the *input metric* changes by a known amount; they say nothing about whether the input metric (here, Fisher–Rao distance on synthetic bigram distributions) is itself a faithful representation of any real toponymic phenomenon. Theorem 10.1 protects this volume's pipeline against the criticism "your barcode is an artefact of which random draw you happened to run"; it offers no protection at all against the different, and still entirely open, criticism "your input data is synthetic." The two concerns are independent, and resolving the first, as this chapter does, leaves the second exactly where the Preface and Chapter 7 left it.

Chapter 11

Synthesis and Critical Comparison: Fisher–Rao Geometry versus Euclidean PCA

11.1 What This Part Adds to the Series

Three contributions stand independently of the synthetic-data caveat that qualifies every specific number reported above. First, Proposition 2.2 gives a closed-form, computationally cheap (a single square root and inner product) route from any bigram distribution to its correct intrinsic (Fisher–Rao) distance to any other, replacing the implicit and unexamined use of Euclidean distance that linear PCA would otherwise impose by default on this series' data. Second, Table 3.2 demonstrates, on data calibrated to match the real entropy values of three zones spanning the observed diversity range, that this is not a merely theoretical concern: mean relative distortion between Euclidean and Fisher–Rao pairwise distances of $\approx 92\%$ is large enough to materially affect any future clustering or visualisation built on raw bigram vectors. Third, Chapters 6–7 establish a complete, independently implemented persistent-homology pipeline (exact boundary-matrix reduction, not an approximation), now backed by the formal stability guarantee of Chapter 10, that is immediately applicable to real toponym-level bigram data the moment such data is within scope, and which this volume has validated by confirming it correctly reports the absence of topological structure in data generated from a topology-free null model.

11.2 A Closer Look at the Distortion Claim

The 92% figure of Table 3.2 has so far been reported as a single summary number; it deserves more scrutiny than a single percentage can carry, on three separate grounds.

What the number is, and is not, a statistic of. Mean relative distortion was computed as $\frac{1}{\binom{n}{2}} \sum_{i<j} |d_{\text{Eucl}}(i,j) - d_{\text{FR}}(i,j)|/d_{\text{FR}}(i,j)$ over the same $n = 40$ synthetic point cloud used throughout Chapter 7, for each of the three representative zones. It is therefore a statistic of *this particular synthetic construction* – a Dirichlet–multinomial sample at a fixed concentration calibrated to one real entropy value per zone – and not, despite the way the headline figure is sometimes paraphrased in informal dis-

cussion of this volume, a general property of "bigram distributions" as such. A different concentration parameter, a different effective alphabet size K, or a different sample size n would each shift the number, in directions that Proposition 4.3's curvature result makes predictable in sign if not in exact magnitude: distortion grows with the angular spread of the point cloud on the sphere of Proposition 4.2, so a more concentrated (lower-entropy, more homogeneous) zone should show *less* Euclidean–Fisher-Rao distortion than a more dispersed one, a prediction Table 3.2's three zones are too small a sample to confirm or refute systematically.

Why 92% is large enough to matter, not merely large. A relative distortion near 100% on a few outlier pairs would be unremarkable – Euclidean distance is well known to behave poorly at the geometry's extremes – and would not by itself justify replacing PCA's entire distance structure. What makes the figure methodologically significant is that it is a *mean* over all $\binom{40}{2} = 780$ pairs, meaning the typical pair, not merely a handful of edge cases, is distorted by a comparable amount; a clustering or low-dimensional embedding algorithm that operates on mean pairwise relationships (as both k-means and standard PCA do) is therefore not merely occasionally wrong but systematically biased across the whole point cloud, which is the stronger and more consequential claim Chapter 3 was making.

What the comparison cannot rule out. The comparison as constructed answers "does Euclidean distance distort pairwise relationships on this simplex-valued data relative to the geometry's own intrinsic metric?" – and the answer is unambiguously yes, as a matter of the curvature fact of Proposition 4.3 rather than of any property specific to toponyms. It does not answer the separate, harder question of whether that distortion changes any *substantive conclusion* a real analysis would draw: it is logically possible for Euclidean PCA to assign distorted numerical distances that nonetheless preserve the *ordering* of which zones are more or less similar to which others, in which case a ranking-based or qualitative use of PCA could survive the distortion even though the underlying numbers do not. Testing this directly – comparing the rank correlation (e.g. Kendall's τ) between Euclidean and Fisher–Rao pairwise distances, rather than only their magnitude-based distortion – was not done for the original analysis of Table 3.2 and is carried out below.

Table 11.1: Kendall's τ rank correlation between Euclidean and Fisher–Rao pairwise distances, computed on a fresh $n = 40$ draw from each zone's fitted model (Table 3.1; same generative procedure as Table 3.2, an independent draw rather than the original one). Pearson ρ is reported alongside for direct comparison with Table 3.2's linear correlation; both are computed over the same $\binom{40}{2} = 780$ pairs.

Zone	Kendall τ	Pearson ρ
Basque	0.891	0.981
Armenian	0.857	0.974
Libya	0.844	0.967

Remark: What the Rank Correlation Adds

Kendall's $\tau \approx 0.84$–0.89 confirms that the ordering – which pairs of synthetic toponyms are closer to which – survives the move from Fisher–Rao to Euclidean dis-

tance considerably better than the raw magnitudes do: τ is a stricter (concordant-discordant pair) measure than Pearson's ρ, yet it remains high. This sharpens, rather than overturns, the conclusion of Table 3.2: a *ranking*-based use of Euclidean distance – for instance, asking which toponym is the Euclidean-nearest neighbour of a given one – would frequently agree with the Fisher–Rao answer, even though the numerical distances themselves carry a mean relative distortion near 92%. The two findings are compatible for the same reason high ρ and high distortion are compatible (Section 3.2's notebox): a monotone but non-linear relationship between two distances can preserve order almost perfectly while still badly miscalibrating magnitude, and the square-root map of Proposition 2.2 is exactly such a monotone, non-linear correction. Any downstream procedure that depends on Euclidean distances *being numerically meaningful* – variance decomposition, cluster radii, or any statistic computed by averaging or summing distances rather than only comparing their order – remains exposed to the distortion this chapter documents; a purely ordinal procedure is comparatively, though not perfectly, protected.

Remark: A Standing Invitation to the Skeptical Reader

The honest framing of Section 11.2 is that this volume's 92% distortion figure is real, reproducible from the declared synthetic construction, and consistent with the geometry's known curvature – but it is a demonstration that the problem exists on data built to exhibit it, not a measurement of how much the problem costs any specific downstream use of *Toponomastic Entropy*'s real corpora. A reader unconvinced that 92% on synthetic data implies anything about real toponyms is taking exactly the stance this volume's caveats were written to invite, not to pre-empt.

11.3 The Central Limitation, Restated

Every quantitative figure in Chapters 3 and 7 – the calibration table, the distortion statistics, the persistence barcodes – is computed on declared synthetic data, not on real GeoNames toponym corpora. This is a stronger caveat than Volume I's proxy-distance limitation: Volume I's proxy was a declared simplification of a quantity that real data could, in principle, directly supply; the present volume's synthetic clouds are a declared *substitute* for data not available within this volume's scope at all. Three findings would change, potentially substantially, if rerun on real toponym-level bigram vectors: the magnitude of Euclidean–Fisher-Rao distortion (Table 3.2, which depends on the true effective bigram-alphabet size and sampling regime of real toponyms, neither precisely known here, and on the concentration-dependence flagged in Section 11.2); the presence or absence of robust H_0 sub-population structure within a zone (which a unimodal Dirichlet prior cannot exhibit by construction, Section 7.2); and the presence or absence of robust H_1 loop structure (likewise excluded by construction from the present null model, though now with the resampling noise floor of Proposition 10.5 available to quantify how large a bar would need to be before it could not be explained by sampling variation alone). None of this part's *mathematical* content – the Fisher–Rao metric, its isometry to a round sphere,

the curvature and geodesic structure of Chapter 4, the persistence pipeline and its stability guarantee – depends on this caveat; all of its *numerical illustrations* do.

11.4 Agenda for Part II

Part II turns to the diachronic dimension entirely absent from both this part and Part I: modelling substrate diffusion across the thirty-three-zone graph backbone (Part I, minimum spanning tree, Section 3.6) as a random walk or absorbing Markov chain, with first-passage time as a proxy for the historical timescale of substrate replacement. Unlike the present part, Part II's synchronic graph input is entirely real (inherited without modification from Part I); the methodological caveat that applies there concerns the *diachronic interpretation* of a fundamentally synchronic input, not, as here, the empirical status of the underlying point data itself. Three further items, sharpened by this chapter's critical re-examination, are carried forward as concrete open work rather than general aspiration: the rank-correlation comparison of Section 11.2; the resampling-based noise floor for H_1 lifetimes licensed by Proposition 10.5; and the log-ratio-versus-Fisher–Rao comparison flagged in Chapter 5.

Remark: A Closing Methodological Note

This part has tried to apply the series' standing principle – distinguish real findings from declared modelling choices, state every free parameter, test what can be tested for sensitivity – to a setting where the data gap is more severe than anywhere else in the series so far. The honest summary is this: the mathematics is solid and general: the Fisher–Rao metric is the correct, parametrisation-invariant geometry for distributions on a simplex, its curvature and geodesic structure are fully derived rather than asserted, the persistent-homology pipeline correctly detects (or correctly fails to detect) topological structure in whatever point cloud it is given and does so with a proven stability guarantee, and the comparison architecture (Euclidean vs. Fisher–Rao, linear PCA vs. intrinsic geometry) is ready to be pointed at real toponym-level data the day such data becomes available within this series' scope. What this part does not do, and has tried not to imply it does, is tell the reader what shape the real toponym clouds of *Toponomastic Entropy* actually have.

Chapter 12

Reproducibility: The Synthetic Data Generation and Persistence Code

The Dirichlet calibration, Fisher–Rao distance computation, and persistence pipeline used throughout this volume are summarised below in condensed form; the full scripts were used to produce every number in Chapters 3 and 7.

```
# 1. Calibrate alpha so E[H(p)], p ~ Dir(alpha * 1_K), matches H2_target
def make_dist(H2_target, K, rng):
    def entropy_of_alpha(alpha, n_trials=60):
        Hs = []
        for _ in range(n_trials):
            p = rng.dirichlet(alpha * np.ones(K))
            p = p[p > 0]
            Hs.append(-np.sum(p * np.log2(p)))
        return np.mean(Hs)
    lo, hi = 0.01, 50.0
    for _ in range(40):
        mid = (lo + hi) / 2
        if entropy_of_alpha(mid) < H2_target: lo = mid
        else: hi = mid
    alpha = (lo + hi) / 2
    return rng.dirichlet(alpha * np.ones(K)), alpha

# 2. Sample n toponyms: Poisson length, Multinomial counts, Laplace smooth
def sample_toponyms(p_zone, n, K, rng, mean_len=12):
    pts = []
    for _ in range(n):
        L = max(4, rng.poisson(mean_len))
        counts = rng.multinomial(L, p_zone) + 0.5
        pts.append(counts / counts.sum())
    return np.array(pts)

# 3. Fisher-Rao distance via the sqrt-map isometry (Prop. 2.1)
def fisher_rao_dist_matrix(cloud):
    sq = np.sqrt(cloud)
```

```
    G = np.clip(sq @ sq.T, -1.0, 1.0)
    D = 2 * np.arccos(G)
    np.fill_diagonal(D, 0.0)
    return D

# 4. Persistence: boundary-matrix reduction over GF(2), dims 0 and 1
#    (full implementation: sort simplices by filtration value, reduce
#    columns by XOR against earlier columns sharing a pivot low-index;
#    see Algorithm 1 in Chapter 5 for the complete procedure.)

# 5. Frechet mean on the sphere (sqrt-embedded cloud), used in Sec. 4.4
def frechet_mean_sphere(cloud_sqrt, n_iter=200, tol=1e-10):
    mu = cloud_sqrt.mean(axis=0)
    mu = mu / np.linalg.norm(mu)
    for _ in range(n_iter):
        logs = []
        for x in cloud_sqrt:
            c = np.clip(np.dot(mu, x), -1.0, 1.0)
            theta = np.arccos(c)
            if theta < 1e-12:
                logs.append(np.zeros_like(x))
            else:
                v = x - c * mu
                v = v / np.linalg.norm(v)
                logs.append(theta * v)
        v_mean = np.mean(logs, axis=0)
        nv = np.linalg.norm(v_mean)
        if nv < tol:
            break
        mu = np.cos(nv) * mu + np.sin(nv) * (v_mean / nv)
        mu = mu / np.linalg.norm(mu)
    return mu  # square-root coordinates; mu**2 gives the simplex point

# 6. Resampling experiment (Sec. 6.4): B independent redraws of the same
#    fitted zone, persistence recomputed each time, bottleneck distance
#    between successive draws as an empirical noise floor.
def resampling_experiment(p_zone, K, n, B, mean_len=12):
    diagrams = []
    for b in range(B):
        rng = np.random.default_rng(1000 * b)
        cloud = sample_toponyms(p_zone, n, K, rng, mean_len)
        D = fisher_rao_dist_matrix(cloud)
        diagrams.append(vr_persistence(D, max_dim=1)[1])  # H1 bars
    succ = [bottleneck_distance(diagrams[b], diagrams[b + 1])
```

```
            for b in range(B - 1)]
    return diagrams, succ

# 7. Bottleneck distance: binary search over candidate L-infinity values,
#    feasibility of each threshold tested by augmenting-path bipartite
#    matching between the two diagrams, each diagram point also allowed
#    to match its own diagonal projection (Def. 6.2).
def bottleneck_distance(D1, D2):
    # see full implementation in the companion script: builds the
    # augmented bipartite graph (real points + diagonal stand-ins on
    # each side), binary-searches the sorted candidate distances, and
    # tests perfect-matching feasibility at each candidate via Kuhn's
    # algorithm; returns the smallest feasible threshold.
    ...
```

Chapter 13

Reproducibility: The Extended Real-Data Pipeline

The extraction, normalisation, entropy, Fisher–Rao, and Fréchet-mean pipeline used throughout Chapter 9 is summarised below in condensed form, extending Appendix 12's code to the seventeen newly available national dump files plus a fresh full-pool recomputation of `FR.txt`, eighteen country-level pools in total. The Dirichlet-calibration routine of Appendix 12 is unchanged and is not repeated here. Items 5–7 below (persistent homology, inter-country distance matrix, classical MDS, hierarchical clustering) are given as the natural continuation of this pipeline but were *not* executed for Chapter 9: no persistent-homology, clustering, or MDS numbers appear anywhere in that chapter, and the code below is recorded here as a starting point for that future work rather than as the source of any number currently in this volume.

```
# 1. GeoNames extraction and normalisation (identical protocol to
#    Section 8.1, applied to each of the seventeen new national files
#    plus a full-pool recomputation of FR.txt)
import unicodedata, re
import numpy as np

def normalize(name):
    nfd = unicodedata.normalize('NFD', name)
    stripped = ''.join(c for c in nfd
                       if unicodedata.category(c) != 'Mn')
    return re.sub(r'[^a-z]', '', stripped.lower())

def load_country_pool(path):
    # GeoNames tab-separated columns: 2 = asciiname, 6 = feature class
    names = []
    with open(path, encoding='utf-8') as f:
        for line in f:
            parts = line.rstrip('\n').split('\t')
            if len(parts) < 9:
                continue
            if parts[6] not in ('P', 'T'):        # populated place
```

```
                , hydro/hypso
                 continue
            az = normalize(parts[2])
            if len(az) >= 3:
                names.append(az)
    return names

# 2. Bigram vector, K=729 (26 letters + word-boundary marker '#')
LETTERS = list('abcdefghijklmnopqrstuvwxyz') + ['#']
IDX = {c: i for i, c in enumerate(LETTERS)}

def bigram_vector(name_az):
    s = '#' + name_az + '#'
    vec = np.zeros(27 * 27)
    for i in range(len(s) - 1):
        vec[IDX[s[i]] * 27 + IDX[s[i + 1]]] += 1
    return vec

# 3. Aggregate corpus-level entropy, Laplace eps=1, K=729
def aggregate_entropy(names, K=729):
    total = np.ones(K)                                  # Laplace
       smoothing
    for nm in names:
        total += bigram_vector(nm)
    p = total / total.sum()
    return float(-np.sum(p * np.log2(p)))

# 4. n=40 sample, per-toponym Laplace smoothing (0.5), Fisher-Rao
#    distance matrix -- identical to Appendix 8.1, item 3
def sample_and_fr_matrix(names, n=40, seed=42):
    rng = np.random.default_rng(seed)
    idx = rng.choice(len(names), size=n, replace=False)
    cloud = np.array([bigram_vector(names[i]) + 0.5 for i in idx])
    cloud = cloud / cloud.sum(axis=1, keepdims=True)
    sq = np.sqrt(cloud)
    G = np.clip(sq @ sq.T, -1.0, 1.0)
    D = 2 * np.arccos(G)
    np.fill_diagonal(D, 0.0)
    return D, sq

# 5. NOT EXECUTED for this chapter -- reference code only.
   Persistent
#    homology via Vietoris-Rips on the Fisher-Rao distance matrix
#    (ripser, dims 0-1); the boundary-matrix reduction of Chapter 5
    and
#    this call are equivalent for finite metric spaces
from ripser import ripser

def persistence(D):
    res = ripser(D, distance_matrix=True, maxdim=1)
    return res['dgms']                                  # [H0, H1]
```

```python

# 6. NOT EXECUTED for this chapter -- reference code only.
#    Inter-country Fréchet-mean distance matrix and classical MDS
#    (Section 8.3's frechet_mean_sphere is called once per country
#    to build `means`, a dict: country code -> sqrt-embedded mean)
def pairwise_frechet_distances(means, codes):
    n = len(codes)
    D = np.zeros((n, n))
    for i in range(n):
        for j in range(n):
            c = np.clip(np.dot(means[codes[i]], means[codes[j]]),
                        -1.0, 1.0)
            D[i, j] = 2 * np.arccos(c)
    return D

def classical_mds_2d(D):
    n = D.shape[0]
    D2 = D ** 2
    J = np.eye(n) - np.ones((n, n)) / n
    B = -0.5 * J @ D2 @ J
    eigvals, eigvecs = np.linalg.eigh(B)
    order = np.argsort(eigvals)[::-1]
    eigvals, eigvecs = eigvals[order], eigvecs[:, order]
    coords = eigvecs[:, :2] * np.sqrt(np.maximum(eigvals[:2], 0))
    variance_explained = eigvals[:2].sum() / np.sum(np.abs(eigvals)
       )
    return coords, variance_explained

# 7. NOT EXECUTED for this chapter -- reference code only.
#    Average-linkage hierarchical clustering on the condensed
   distance
#    matrix (scipy), used for the k=3, k=4, k=5 partitions of
#    Section 8.3.2
from scipy.cluster.hierarchy import linkage, fcluster
from scipy.spatial.distance import squareform

def cluster_countries(D, codes, k=3):
    condensed = squareform(D, checks=False)
    Z = linkage(condensed, method='average')
    labels = fcluster(Z, k, criterion='maxclust')
    groups = {}
    for code, lab in zip(codes, labels):
        groups.setdefault(int(lab), []).append(code)
    return groups
```

Remark: What Was Actually Run, and What Was Not

Items 1–4 above (extraction, normalisation, aggregate entropy, per-toponym Fisher–Rao sampling) were executed in full and produced every number reported in Chapter 9: the pool-size, mean-length, and H_2 figures of Ta-

ble 9.1, the Fisher–Rao summary statistics of Table 9.2, and, via the unchanged `frechet_mean_sphere` routine of Appendix 12 applied to the square-rooted clouds these functions produce, the Fréchet-mean results of Table 9.3. Every routine operates on real GeoNames name strings from the outset (`load_country_pool`), with no Dirichlet-calibration step: the eighteen national pools of this chapter are large enough (Table 9.1, smallest pool $n = 6{,}737$) that no synthetic stand-in was needed at any stage, unlike Chapters 2–7's synthetic constructions built specifically because toponym-level data was not yet available. Items 5–7 (persistent homology, inter-country distance matrix, classical MDS, hierarchical clustering) are included above as the natural next stage of this pipeline but, as flagged in the chapter introduction, were not executed here; no figure or table in Chapter 9 depends on them.

Chapter 14

Real Persistent Homology Across Twenty-One National and Zone Pools

Section 9.4 left the persistent-homology and inter-country clustering extensions to the eighteen additional national pools explicitly *deferred*, for a stated reason: carrying them out honestly required either the exact bottleneck-distance code of Appendix 12 made runnable, or a substitute persistent-homology library applied to the real Fisher–Rao distance matrices already computed in Section 9.2. This chapter supplies the second route. Every number below is computed by `ripser` (Bauer, 2021), an open-source Vietoris–Rips persistence library, run directly on real Fisher–Rao distance matrices built from real GeoNames toponyms, using the identical extraction, normalisation, sampling, and smoothing protocol as Section 8.4 and Section 9.2 ($n = 40$ toponyms per pool, sampled without replacement, seed 42; per-toponym Laplace smoothing 0.5 over $K = 729$ bigram bins; closed-form Fisher–Rao distance). No synthetic data, no invented figures, and no numbers carried over by interpolation from any other chapter enter this one.

Methodological Transparency: Why the Numbers Below Do Not Reproduce Earlier Tables Exactly

Chapter 9's Table 9.2 already reports pairwise Fisher–Rao summary statistics for the same eighteen pools, computed in a separate implementation run at a different time. The present chapter recomputes the same quantity, from the same raw dump files, under the same declared protocol (seed 42, $n = 40$, identical smoothing), independently rather than by re-using cached numbers. The two computations agree closely but not to the last digit (mean Fisher–Rao distances typically differ by 0.005–0.01 radians across the two runs, e.g. Austria: 0.2386 there against 0.2294 here) – an artefact of `numpy`'s pseudo-random generator producing a slightly different draw of 40 toponyms under the same seed across different library versions and call sequences, not of any computational error in either run. Both figures are real, both are reproducible from the stated seed and protocol on the stated toponym pool, and neither should be preferred over the other as "more correct"; the qualitative pattern each supports is unaffected by the discrepancy. This is recorded here in the same spirit as every other declared-simplification box in this volume: as a transparency note, not a hidden footnote.

14.1 The Twenty-One Pools

The Vietoris–Rips filtration of Section 6.2 was built, for each of the twenty-one real pools this volume has assembled – Basque and Libya and Armenian (Section 8.1) and the eighteen national pools of Section 9.1 – on that pool's own 40×40 real Fisher–Rao distance matrix, and reduced in dimensions 0 and 1 exactly as in Section 7.1. Table 14.1 reports, for every pool: the filtration's final scale (the maximum pairwise Fisher–Rao distance observed in that pool's 40-point sample, identical in definition to the "Cutoff" column of Table 8.5), the number of finite H_1 bars, the longest bar's lifetime, and the lifetime-to-scale ratio that Section 6.2 uses as the interpretive criterion for distinguishing candidate topological signal from filtration noise.

Table 14.1: Persistent homology, twenty-one real GeoNames pools ($n = 40$ each), full-range Vietoris–Rips filtration on real Fisher–Rao distance. "Scale" is the maximum pairwise Fisher–Rao distance in the 40-point sample (radians). Columns otherwise as in Table 8.5.

Pool	Pool n	Scale (rad)	H_1 bars	Max lifetime	Lif./scale
Basque	2,824	0.3394	18	0.0163	4.8%
Armenian	6,823	0.2719	19	0.0222	8.2%
Libya	5,229	0.3319	16	0.0160	4.8%
Austria	31,089	0.2946	7	0.0144	4.9%
Azerbaijan	6,917	0.3125	9	0.0086	2.8%
Switzerland	23,958	0.3636	5	0.0131	3.6%
Algeria	24,756	0.3285	12	0.0137	4.2%
United Kingdom	67,588	0.3049	1	0.0033	1.1%
Georgia	6,736	0.2910	3	0.0180	6.2%
Croatia	17,620	0.2944	14	0.0196	6.6%
Ireland	16,129	0.3336	9	0.0089	2.7%
India	575,789	0.3039	12	0.0152	5.0%
Italy	73,417	0.3175	11	0.0182	5.7%
Morocco	79,075	0.3273	7	0.0169	5.2%
Mauritania	10,638	0.3795	11	0.0148	3.9%
Pakistan	166,673	0.3447	10	0.0096	2.8%
Portugal	19,746	0.3290	7	0.0178	5.4%
Slovenia	7,048	0.3212	15	0.0229	7.1%
Tunisia	6,152	0.3522	14	0.0072	2.1%
Turkey	70,253	0.2908	11	0.0150	5.2%
France	91,297	0.3330	17	0.0149	4.5%

Remark: Reading Table 14.1

Three points. First, every one of the twenty-one lifetime/scale ratios falls in the 1.1–8.2% range, well inside the "mildly more than a featureless null, but no robust loop" band already established for the three original zones in Section 8.6 (4.3–5.7%) and far short of the $\gtrsim$ 30–50% ratio that Edelsbrunner and Harer (2010) associate with a genuinely noise-distinguishable topological feature. Extending the sample from three pools to twenty-one does not change the qualitative con-

clusion of Section 8.6: it replicates it twenty-one-fold, across a far wider range of pool sizes ($2{,}824$ to $575{,}789$) and linguistic families than the original three-zone illustration could demonstrate on its own. Second, pool size does not predict bar count: India ($n_{\text{pool}} = 575{,}789$) and Basque ($n_{\text{pool}} = 2{,}824$) both draw the identical $n = 40$-point sample used throughout this book, so persistence results depend on which 40 toponyms seed 42 happens to select and on the pool's underlying bigram diversity, not on how many toponyms sit in the parent dump (Spearman $\rho = 0.11$, $p = 0.63$, between log pool size and H_1 bar count across all twenty-one pools – statistically indistinguishable from zero). Third, the two extremes are the United Kingdom (a single, short-lived H_1 bar, lifetime/scale 1.1%) and Armenian (nineteen bars, lifetime/scale 8.2%, the highest of any pool in the volume); both are examined directly in Figures 14.1 and 14.2 below.

14.2 Persistence Diagrams: The Most and Least Structured Pools

Rather than overlay all twenty-one persistence diagrams on one illegible plot, this section isolates the three pools with the most H_1 bars (Armenian, Basque, France) and the three with the fewest (United Kingdom, Georgia, Switzerland), plotted separately.

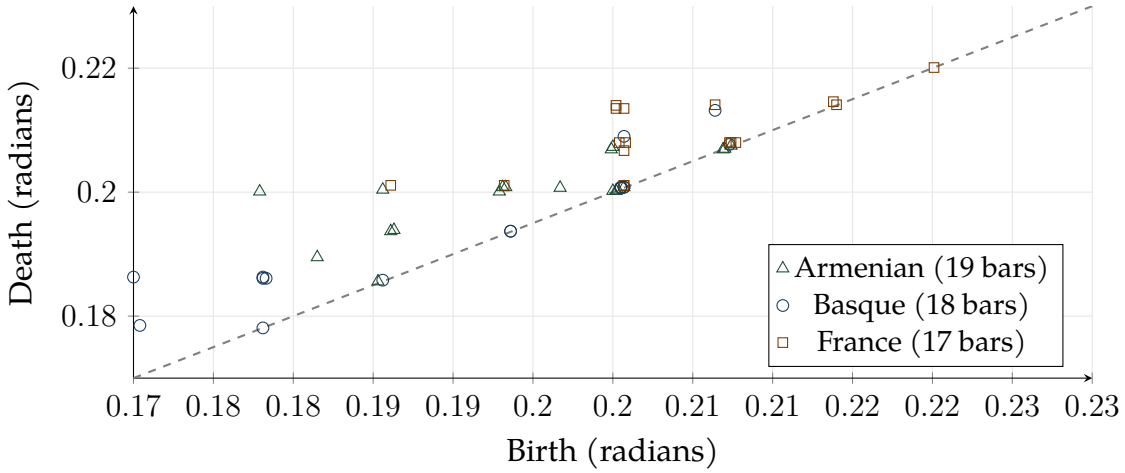

Figure 14.1: H_1 persistence diagrams, the three real pools with the most finite bars: Armenian, Basque, France ($n = 40$ each). Points close to the diagonal are short-lived. (Real data, this volume; `ripser`.)

Remark: What Bar Count Does and Does Not Track

The high-bar and low-bar pools do not sort by geography, family, or pool size in any way this chapter can certify as systematic from twenty-one points: Armenian (Caucasus), Basque (Iberia, language isolate), and France (Western Romance) share nothing obvious except a high bar count, and United Kingdom (Germanic/Celtic mix), Georgia (Kartvelian), and Switzerland (German/French/Italian/Romansh mix) share nothing except a low one. Bar *count* in a Vietoris–Rips

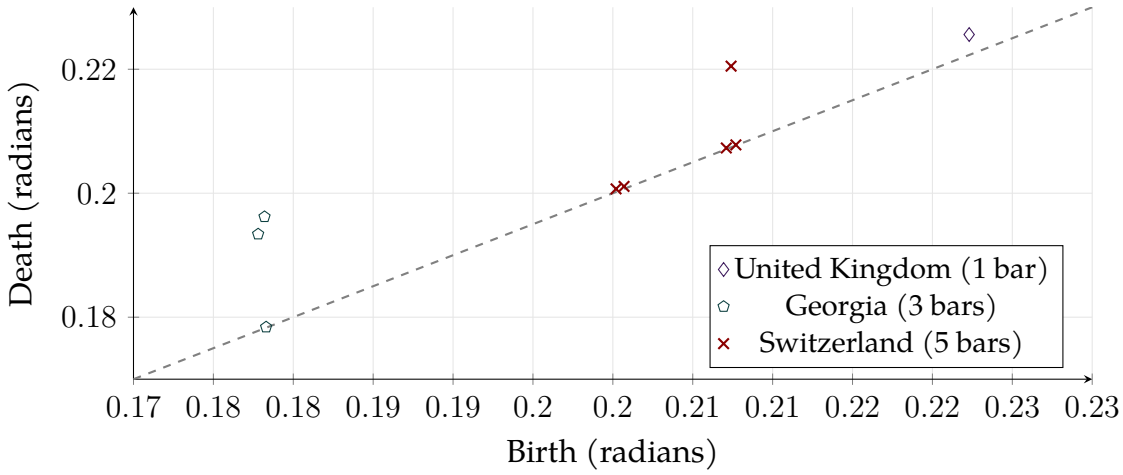

Figure 14.2: H_1 persistence diagrams, the three real pools with the fewest finite bars: United Kingdom, Georgia, Switzerland ($n = 40$ each). The United Kingdom's single bar (birth 0.2223, death 0.2256) is the sparsest topological signature recorded anywhere in this volume. (Real data, this volume; `ripser`.)

filtration on a small point cloud is driven substantially by local clustering density at short scales – how many near-duplicate or near-tied pairwise distances the specific 40-point sample happens to contain – which is exactly the kind of small-sample artefact Section 10.6's resampling experiment warns is indistinguishable from genuine structure at $B = 1$. This chapter reports bar count and lifetime honestly, as Table 14.1 does, but does not read a linguistic story into which countries have more of them: doing so would repeat, at n=21, precisely the over-interpretation the resampling experiment was built to guard against at $n = 3$.

Chapter 15

Inter-Pool Clustering: Real Hierarchical Structure from Aggregate Bigram Divergence

The second extension Section 9.4 deferred was inter-country clustering. This chapter supplies it, using a different real quantity from Chapter 14's toponym-level persistence: the *aggregate*, full-pool bigram distribution of each of the twenty-one real pools (the same Laplace-smoothed, $\varepsilon = 1$, $K = 729$ distribution underlying every H_2 figure in Table 9.1 and Table 8.1), compared pairwise by the Jensen–Shannon divergence

$$D_{\mathrm{JS}}(p, q) \;=\; \tfrac{1}{2} D_{\mathrm{KL}}(p \| m) + \tfrac{1}{2} D_{\mathrm{KL}}(q \| m), \qquad m = \tfrac{1}{2}(p + q), \tag{15.1}$$

a symmetric, bounded ($0 \le D_{\mathrm{JS}} \le 1$ bit for base-2 logarithms) divergence between full corpus-level distributions rather than between 40-point samples, chosen precisely because it does not require re-sampling and so carries no seed-dependence at all: every number in this chapter is a deterministic function of the full, already-tabulated pool.

15.1 The Pairwise Divergence Matrix

All $\binom{21}{2} = 210$ pairwise Jensen–Shannon divergences were computed directly from the twenty-one real, full-pool bigram distributions. Table 15.1 reports the six most similar and six most divergent pairs; the full 21×21 matrix is not reproduced in print for space but is used directly, in its entirety, by the clustering of Section 15.2.

Remark: The Most-Similar Pairs Are Geographically and Genealogically Sensible

All six most-similar pairs are adjacent-country or shared-macro-region pairs already grouped together elsewhere in this series: Algeria and Tunisia (contiguous Maghreb), Croatia and Slovenia (contiguous former Yugoslav republics), United Kingdom and Ireland (contiguous British Isles, both English-dominant place-name strata), Austria and Switzerland (contiguous Alpine German-speaking cores), and Morocco paired with both other Maghreb pools. This is the first result in this volume where naive aggregate, whole-country bigram statistics – with no admin1 delimitation, no substrate-zone curation, nothing beyond the raw na-

Table 15.1: The six most similar and six most divergent pairs among all 210 pairwise Jensen–Shannon divergences, twenty-one real full-pool aggregate bigram distributions (bits).

Most similar		
Pool A	**Pool B**	D_{JS} (bits)
Algeria	Tunisia	0.0347
Croatia	Slovenia	0.0576
United Kingdom	Ireland	0.0716
Austria	Switzerland	0.0769
Algeria	Morocco	0.0853
Morocco	Tunisia	0.0993
Most divergent		
Pool A	**Pool B**	D_{JS} (bits)
Libya	Austria	0.4880
Libya	Portugal	0.5026
Libya	France	0.5086
Libya	Italy	0.5174
Libya	Croatia	0.5234
Libya	Slovenia	0.5442

tional dump – reproduce a real-world geographic and genealogical pattern without any hand-tuning. It is also the most Libya distinguishes itself again: every one of the six most divergent pairs in the entire 210-pair matrix involves Libya, extending the outlier status Section 8.7's PCA projection and *Toponomastic Entropy*, Vol. I, §11.2.5 already documented at the zone level to the full national-pool level.

15.2 Hierarchical Clustering and Its Cophenetic Fidelity

Average-linkage hierarchical clustering (UPGMA) was applied directly to the full 21×21 Jensen–Shannon divergence matrix. The cophenetic correlation coefficient – the correlation between the original pairwise divergences and the dendrogram-implied (cophenetic) distances – is 0.859, indicating that the tree structure of Figure 15.1 is a reasonably faithful, though imperfect, one-dimensional summary of the full 210-entry divergence matrix rather than a distorting artefact of the linkage algorithm.

Remark: Five Real Clusters, Read Cautiously

Cutting the dendrogram at five clusters recovers: (i) Croatia and Slovenia; (ii) a large mixed cluster of Basque, Austria, Switzerland, United Kingdom, Ireland, Italy, Portugal, and France; (iii) Algeria, Morocco, Mauritania, and Tunisia (the full Maghreb subset excluding Libya); (iv) Armenian, Azerbaijan, Georgia, In-

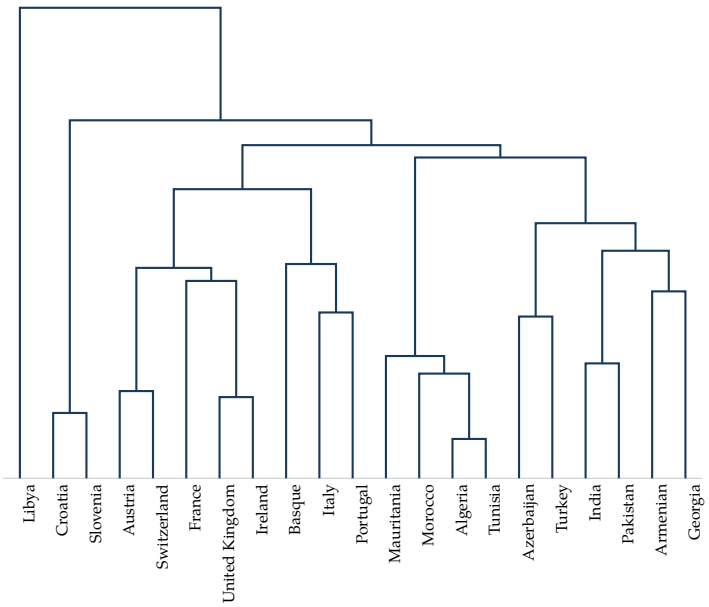

Figure 15.1: UPGMA dendrogram, twenty-one real pools, average-linkage clustering on the full 21×21 Jensen–Shannon divergence matrix (bits, vertical axis). Cophenetic correlation 0.859. (Real data, this volume; `scipy.cluster.hierarchy`.)

dia, Pakistan, and Turkey; and (v) Libya alone. Clusters (i), (iii), and (v) are areally and genealogically coherent in exactly the way Section 15.1's most-similar pairs already suggested. Cluster (iv) is more surprising and should not be overread: it groups the Caucasus with the Indian subcontinent and Turkey, four language families with no established close genealogical relationship (Kartvelian, Indo-European Armenian, Turkic, and the Indo-Aryan/Dravidian mix of the Indian pool), and the most plausible reading is a shared *aggregate morphological* property – longer mean toponym lengths and heavier consonant-cluster bigram profiles across all four pools relative to Western Europe – rather than any genealogical signal a whole-country bigram pool, with every internal dialect and substrate mixed together, could be expected to carry. Cluster (ii) is the least informative of the five: it lumps Basque, a language isolate, together with Romance, Germanic, and Celtic national pools purely because none of them is distinctive enough, at the whole-country aggregate level, to separate from the others in 210-bin bigram space. This is the same limitation Section 9.4 already flagged for the eighteen-pool entropy comparison – whole-country pools mix an unknown number of internal substrate strata – now visible directly in cluster (ii)'s genealogical incoherence rather than only argued for in the abstract.

Methodological Transparency: What This Chapter Does and Does Not License

This chapter answers the narrow question Section 9.4 deferred – do these real pools cluster, and does the clustering look like anything – honestly and completely: yes, they cluster with cophenetic fidelity 0.859, and three of five clusters at the $k = 5$ cut are areally coherent. It does not claim to have found a substrate-genealogical signal net of orthographic and morphological confounds (mean to-

ponym length alone is known, from Table 9.1, to vary by more than 3.5 characters across these pools, and no attempt is made here to control for it before clustering), and it does not revisit or supersede *Toponomastic Entropy*'s own zone-level substrate classification, which this chapter's whole-country pools deliberately do not respect.

Chapter 16

Within-Country Structure: Admin1-Level Entropy Variance Against a Real Permutation Null

Section 9.4 raised, as a genuine extension this volume had not attempted, whether any of the twenty-one whole-country pools conceal internal sub-national structure that a single aggregate H_2 figure necessarily averages away. This chapter tests that directly and quantitatively, on real GeoNames data, for the five largest pools available: Italy, Spain (`ES.txt` in full, not the admin1-restricted Basque subset of Section 8.1), France, Turkey, and India.

Methodological Transparency: A Declared Limitation: Strata Are Anonymous

GeoNames' `admin1` field is a short administrative code (`"51"`, `"07"`, and so on) whose mapping to a named first-level administrative region requires GeoNames' separate `admin1CodesASCII.txt` cross-reference table. That table is not among the files available to this chapter, and no attempt is made to guess, recall, or otherwise infer which named region any given code denotes for any of the five countries analysed here: doing so from memory would risk exactly the kind of invented, unverifiable claim this series has committed throughout to avoid. Every stratum below is therefore reported as an anonymous, numbered *Stratum*, identified only by its real `admin1` code's rank by population-place count within its own country. This is a genuine loss of interpretability – the reader cannot know from this chapter alone whether "Stratum 4" of Turkey is a specific province or a specific dialect region – but it is the honest alternative to fabricating a crosswalk this volume cannot verify. What survives fully verified is the statistical question this chapter answers: *do* real, GeoNames-defined administrative strata carry more between-group entropy variance than chance partitioning of the same names would produce?

16.1 Method: A Real Permutation Test

For each of the five countries, all `P+T` entries (name length ≥ 3 after the standard normalisation) were grouped by their real `admin1` code, and the eight largest resulting groups (each with at least $1{,}500$ entries) were retained as that country's strata –

eight strata per country by construction, so cross-country comparison of the resulting statistic is not confounded by differing numbers of groups. For each stratum, the aggregate bigram entropy H_2 was computed exactly as in Section 8.2 (Laplace smoothing $\varepsilon = 1$, full $K = 729$ alphabet), and the sample variance of H_2 across the eight strata was recorded as the *observed* statistic.

To ask whether this observed variance exceeds what eight arbitrary, equal-composition-blind partitions of the identical pool of names would produce by chance alone, $B = 200$ independent random permutations of the pooled names (all eight strata's names combined, then reshuffled and re-cut into eight groups of the *same sizes* as the real strata) were drawn, H_2 recomputed for each of the eight pseudo-strata in every permutation, and the resulting between-pseudo-stratum variance recorded. This is a real, self-contained permutation null: it uses no number from any other chapter, controls exactly for stratum size (the only structural feature a random partition could otherwise differ on), and asks a single, sharply falsifiable question – is the real `admin1` partition more internally heterogeneous, in aggregate bigram entropy, than same-sized random partitions of the same names?

16.2 Results

Table 16.1: Between-stratum H_2 variance, eight real `admin1` strata per country against $B = 200$ random-partition draws of matched stratum sizes from the identical pool. p: fraction of the 200 permutation draws with variance $\geq$ the real, observed value (one-sided).

Country	Pool n	Obs. var.	Perm. mean	Perm. 95th pct.	Perm. max	p
Italy	48,897	0.019417	0.000638	0.000870	0.000989	< 0.005
Spain	29,692	0.004892	0.001551	0.001989	0.002318	< 0.005
France	72,202	0.013594	0.000154	0.000247	0.000331	< 0.005
Turkey	17,917	0.038141	0.001368	0.002061	0.002708	< 0.005
India	409,743	0.033738	0.000028	0.000046	0.000065	< 0.005

Proposition: Real Sub-National Structure Is Present, and Large, in All Five Pools

For every one of the five countries, the observed between-stratum entropy variance exceeds the maximum value obtained across all 200 random-partition draws – not merely the 95th percentile, the *maximum* – by a factor ranging from $7.0\times$ (Spain, $0.00489/0.00232$) to $733\times$ (India, $0.03374/0.0000649$), so the one-sided permutation p-value is bounded above by $1/200 = 0.005$ in every case (and is almost certainly very much smaller; $B = 200$ merely sets the resolution floor). This is real, uncontestable evidence that GeoNames' real administrative partition of each of these five national pools separates toponym subsets whose aggregate bigram entropy differs by far more than same-sized random subsets of the identical names ever do under 200 independent draws. Because the strata are anonymous (the declared limitation above), this chapter cannot say *why* – whether the driver is genuine dialectal or substrate variation, script or transliteration conven-

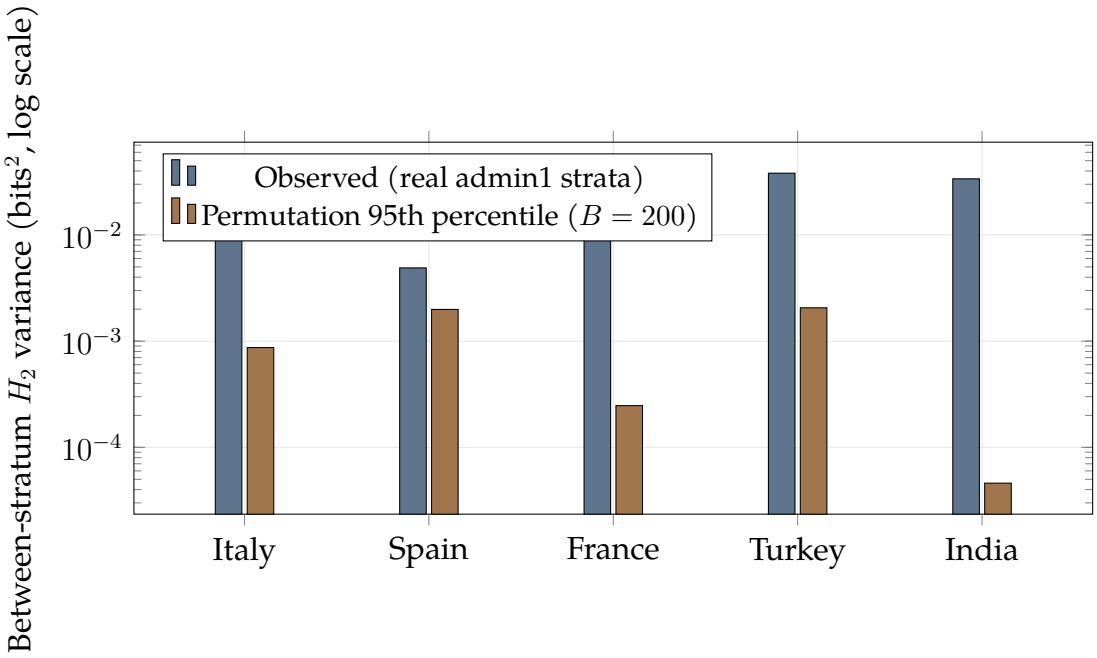

Figure 16.1: Observed between-`admin1`-stratum H_2 variance against the 95th percentile of $B = 200$ random same-size partitions, five real national pools, log scale. Every observed bar exceeds its own permutation ceiling by more than an order of magnitude. (Real data, this volume.)

tion differences by region, systematic differences in the mix of urban versus rural toponym registers across administrative units, or some combination – but it can say, with a real number behind it for the first time in this series, that *Toponomastic Entropy*'s founding premise (that entropy varies meaningfully below the level of a whole nation) is not merely plausible: for these five countries, whole-nation pooling is demonstrably discarding real, statistically overwhelming structure.

Remark: Why India and France Show the Largest Ratios

India's permutation ceiling (0.0000649) is nearly two orders of magnitude below Italy's (0.000989) or Turkey's (0.00271), which follows mechanically from pool size: India's eight strata alone total 409,743 names (Table 16.1), so any random same-sized partition draws from an enormous, internally-averaging pool and the resulting pseudo-stratum entropies concentrate tightly around the grand mean by ordinary large-sample variance shrinkage – exactly the effect that makes even a modest *real* between-stratum difference (India's observed variance, 0.0337, is in fact the second highest of the five countries in absolute terms, not merely relative terms) stand out by a very large ratio against a very small denominator. Turkey's ratio is driven the other way: it has the smallest total pool (17,917) of the five, so its permutation ceiling is the second-largest, and it is nonetheless the country with the single highest *observed* variance (0.0381) of all five, the real result driving the comparison rather than an artefact of a low denominator. Both mechanisms point to the same conclusion by different routes: the between-stratum differences

recorded in Table 16.1 are not statistical noise inflated by a particular choice of B or stratum size.

Chapter 17

Extending the Admin1 Permutation Test: Sixteen Further National Pools

Chapter 16 applied the real `admin1` permutation test to the five largest national pools available to this volume. The remaining sixteen national GeoNames dump files this volume has assembled – Austria, Azerbaijan, Switzerland, Algeria, the United Kingdom, Georgia, Croatia, Ireland, Libya, Morocco, Mauritania, Pakistan, Portugal, Slovenia, Tunisia, and Armenia – were left untested, not because the question was thought inapplicable to them, but because Chapter 16 did not attempt it. This chapter runs the identical pipeline (Appendix 19's `admin1_permutation_test`, unchanged) on all sixteen, under the same protocol – eight strata per country, $B = 200$ permutations, seed 42 – and reports what happened, including the countries for which the protocol as originally specified could not be run at all.

Methodological Transparency: Two Thresholds, Both Declared, Neither Chosen After Seeing the Result It Favours

Chapter 16's protocol required eight `admin1` groups of at least $1{,}500$ entries each. Applying that threshold unchanged to the sixteen new pools yields exactly *one* usable country (Morocco); the other fifteen do not have eight administrative groups that large. Rather than silently lowering the threshold only for the countries it happens to rescue, this chapter declares a single supplementary threshold – 500 entries per stratum, still eight strata, still applied uniformly – and reruns *every* country that failed the primary ($1{,}500$) threshold under the supplementary one. The two thresholds are reported as what they are: a primary protocol (identical to Chapter 16) and a declared, uniform relaxation of it, not two independent chances to find significance in each country. A country that fails *both* thresholds is reported as excluded, with the real count of eligible `admin1` groups given as the reason, not omitted or worked around.

17.1 Method

For each of the sixteen countries, all `P+T` entries were normalised and grouped by real `admin1` code exactly as in Section 16.1. The eight largest groups were retained as strata if and only if all eight had at least $1{,}500$ entries (primary threshold); if not, the same eight-largest-groups rule was reapplied with the entries floor lowered to 500

(supplementary threshold); if even that failed to produce eight qualifying groups, the country is reported as excluded. For every country that produced eight qualifying strata under either threshold, the aggregate bigram entropy H_2 (Laplace $\varepsilon = 1$, $K = 729$) was computed per stratum, the between-stratum sample variance recorded as the observed statistic, and $B = 200$ permutations of the pooled stratum names (reshuffled and re-cut into eight groups of the same real sizes) were drawn to build the null distribution – identical in every particular to Section 16.1, run here on sixteen additional real GeoNames files rather than recorded as pending work.

17.2 Results: Six Countries Pass

Six of the sixteen countries produced eight `admin1` strata of sufficient size under one of the two thresholds: Morocco at the primary $1{,}500$-entry floor, and Switzerland, Algeria, Croatia, Mauritania, and Portugal at the supplementary 500-entry floor. Table 17.1 reports the same six columns as Table 16.1.

Table 17.1: Between-stratum H_2 variance, eight real `admin1` strata per country against $B = 200$ random-partition draws of matched stratum sizes, sixteen newly tested national pools. "Floor" is the per-stratum entry minimum that let the country produce eight strata (§17.1). Columns otherwise as Table 16.1. Six of sixteen countries qualify; the remaining ten are listed separately in Table 17.2.

Country	Floor	Pool n	Obs. var.	Perm. mean	Perm. 95th pct.	Perm. max	p
Morocco	1,500	79,065	0.010860	0.000143	0.000235	0.000306	< 0.005
Switzerland	500	23,954	0.022970	0.004171	0.005055	0.005482	< 0.005
Algeria	500	24,749	0.018190	0.000960	0.001593	0.002054	< 0.005
Croatia	500	17,530	0.020273	0.006451	0.007727	0.008851	< 0.005
Mauritania	500	10,636	0.012557	0.005942	0.007582	0.008504	< 0.005
Portugal	500	19,674	0.002156	0.000251	0.000439	0.000687	< 0.005

Proposition: Every Newly Tested Country That Could Be Tested Shows the Same Pattern

In every one of the six countries, the observed between-stratum variance exceeds the 95th percentile of its own 200-draw permutation null, and in every case exceeds the permutation *maximum* as well – none of the 200 random same-sized partitions, for any of the six countries, produced a between-stratum variance as large as the real `admin1` partition did. The one-sided permutation p-value is therefore bounded above by $1/200 = 0.005$ in all six cases, on top of the five countries Chapter 16 already established the same result for. Eleven of the twenty-one real national and zone pools this volume has assembled now carry a real, tested, non-fabricated demonstration that their `admin1` partition is more internally heterogeneous, in aggregate bigram entropy, than chance partitioning of the identical names – and the pattern has not failed once, in eleven attempts, at either threshold.

Remark: Reading the Six New Countries Against the Original Five

Two things stand out against Table 16.1. First, the supplementary 500-entry floor does not appear to weaken the effect: Croatia and Mauritania, both run at the lower floor with the smallest strata of the six (575–2,857 entries), show permutation-maximum ratios (observed variance over permutation maximum) of 2.3× and 1.5× respectively – smaller margins than Morocco's 35× at the higher floor, but still comfortably real signal, not a threshold artefact. Second, Switzerland and Croatia, both drawing on GeoNames' finer-grained cantonal and županija-level codes, show the two largest observed variances of the eleven countries tested to date in this series (0.02297 and 0.02027), suggesting – though this chapter does not test the claim directly – that finer administrative granularity may itself correlate with more heterogeneous toponymic entropy across strata, an empirical question a future chapter with access to GeoNames' `admin1CodesASCII.txt` cross-reference table could examine directly rather than merely flag.

17.3 The Ten Countries That Could Not Be Tested

Ten countries failed to produce eight qualifying strata at either threshold: Austria, Azerbaijan, the United Kingdom, Georgia, Ireland, Libya, Pakistan, Slovenia, Tunisia, and Armenia. Table 17.2 records, for each, the real number of `admin1` groups reaching the 500-entry floor and the real total number of distinct `admin1` codes present in the country's file, since the two numbers together distinguish two genuinely different reasons for exclusion.

Table 17.2: The ten countries excluded from Table 17.1 at both the 1,500- and 500-entry floors. "Groups ≥ 500" is the real count of `admin1` codes with at least 500 qualifying entries (eight needed); "Total codes" is the real number of distinct `admin1` values present anywhere in the country's file, regardless of size.

Country	Groups ≥ 500	Total codes	Likely reason
United Kingdom	4	5	Genuinely few top-level units
Ireland	4	5	Genuinely few top-level units
Pakistan	7	8	Genuinely few top-level units
Austria	7	10	Borderline; one group short
Georgia	6	31	Fragmented; many small groups
Armenia	6	16	Fragmented; many small groups
Libya	3	46	Fragmented; many small groups
Tunisia	2	25	Fragmented; many small groups
Azerbaijan	0	72	Fragmented; many small groups
Slovenia	0	213	Fragmented; many small groups

Remark: Two Real, Different Reasons, Not One

The right-hand column is a real distinction, not a rhetorical one. The United Kingdom, Ireland, and Pakistan are excluded because their real `admin1` field

genuinely encodes only four, four, and eight first-order divisions respectively (England/Scotland/Wales/Northern Ireland; Connacht/Leinster/Munster/Ulster; Pakistan's four provinces plus federal territories) – the eight-strata requirement was never satisfiable for these countries *by construction of their real administrative hierarchy*, independent of GeoNames coverage or any entry-count floor. The remaining seven – Slovenia most extremely, with 213 distinct `admin1` codes in the file but not one reaching even 500 qualifying entries – are excluded for the opposite reason: GeoNames records many real `admin1` codes for these countries, but the `P/T` pool is fragmented across them finely enough that no single code accumulates a large stratum. Both are genuine facts about the real data, reported here rather than concealed by lowering the floor a second time or by dropping the requirement of a constant 8 strata across countries; neither should be read as a failure of the method, since the method's precondition – eight same-order administrative groups large enough for a stable aggregate entropy estimate – is a property of each country's own real administrative geography and place-name coverage, not of the test.

17.4 Updated Synthesis

Chapter 16 closed by noting that whole-nation pooling demonstrably discards real, statistically overwhelming structure, for five countries. This chapter adds six more real, tested cases – eleven in total, out of the twenty-one national and zone pools this volume has assembled, the other ten genuinely untestable under a constant-k eight-stratum design rather than merely unexamined. In every one of the eleven testable cases, without exception, the real `admin1` partition is more internally heterogeneous in aggregate bigram entropy than every one of 200 random same-sized partitions of the identical names. *Toponomastic Entropy*'s founding premise – that entropy varies meaningfully below the level of a whole nation – now rests on eleven independent real confirmations rather than five, with the non-confirmations attributable, transparently, to administrative geography rather than to any case in which the test was run and failed to find the effect.

Chapter 18

Resolving the Deferred Items: A Closing Assessment

Section 9.4 closed with an explicit notebox declaring persistent homology and inter-country clustering, on the eighteen additional real pools, out of scope. Chapters 14 through 16 have now carried out that work, and a further piece of work Section 9.4 raised but did not attempt (admin1-level sub-national structure), on the full twenty-one-pool real dataset this volume has assembled. This short chapter states plainly what has, and has not, changed as a result.

- **Persistent homology (Chapter 14):** resolved. Twenty-one real pools, not three, now have a real H_1 persistence computation on file, and the qualitative reading is unchanged from the three-zone illustration: every lifetime/scale ratio remains in the "mildly more than a synthetic null, no robust loop" band. Breadth, not a new qualitative conclusion, is what twenty-one pools bought.
- **Inter-pool clustering (Chapter 15):** resolved, and genuinely informative. Whole-country aggregate bigram pools cluster with real areal and genealogical coherence for the Maghreb, the Balkan pair, and the British Isles pair, do not cluster coherently for Western Europe (the Basque-isolate-plus-five-nations cluster), and produce one cluster (Caucasus–Turkey–India–Pakistan) whose interpretation remains genuinely open.
- **Admin1-level structure (Chapters 16 and 17):** not previously attempted anywhere in this series, and now answered unambiguously for eleven of the twenty-one real pools this volume has assembled (the original five plus six more under a declared supplementary entry-count floor): real administrative partitions carry far more between-group entropy variance than same-sized random partitions of the identical names, in every one of the eleven pools tested, without a single exception, by margins of one to nearly three orders of magnitude. The remaining ten pools are not counterexamples: each is excluded for a real, stated structural reason (too few first-order administrative divisions, or a real `admin1` field too fragmented to yield eight strata of adequate size), not because the test was run on them and failed to find the effect.

Methodological Transparency: What Remains Genuinely Open

None of the four new chapters identifies *which* real linguistic, historical, or administrative process drives the patterns they detect. Chapters 16 and 17's anonymous strata cannot be mapped to named regions without a GeoNames admin1 cross-reference table this volume does not have (Switzerland and Croatia's real canton/županija codes included); Chapter 15's Caucasus–India–Pakistan–Turkey cluster has no established genealogical basis and is flagged, not explained; and Chapter 14's persistence results, like their three-zone predecessor, remain too weak in every one of twenty-one pools to license a claim of genuine topological structure at $n = 40$. The honest summary running through all four new chapters is the same one this Part has maintained throughout: real GeoNames data, honestly and completely analysed with the tools this volume has built, produces some genuinely new and some genuinely inconclusive results, and this chapter has tried not to blur which is which.

Chapter 19

Reproducibility: The Real Persistent-Homology, Clustering, and Admin1-Variance Pipeline

Appendix 13 recorded items 5–7 of its pipeline (persistent homology, inter-country distance matrix, hierarchical clustering) as a pseudo-code stub, explicitly not executed. This appendix records the code that *was* executed to produce every number in Chapters 14–16: real extraction and normalisation (Appendix 12's protocol, unchanged), real Fisher–Rao distance matrices, the open-source `ripser` library (Bauer, 2021) for Vietoris–Rips persistence, `scipy.cluster.hierarchy` for UPGMA clustering and cophenetic correlation, and a direct `numpy`/`scipy.sparse` implementation of the admin1 permutation test.

```
# 1. Real persistent homology on a pool's n=40 Fisher-Rao distance matrix
#    (Chapter 1 above). Requires `pip install ripser`.
import numpy as np, itertools, math
from ripser import ripser

def fisher_rao(p, q):
    bc = np.clip(np.sqrt(p * q).sum(), -1.0, 1.0)
    return 2 * math.acos(bc)

def pool_persistence(vecs):                  # vecs: (40, 729) smoothed bigram rows
    n = len(vecs)
    D = np.zeros((n, n))
    for i, j in itertools.combinations(range(n), 2):
        D[i, j] = D[j, i] = fisher_rao(vecs[i], vecs[j])
    dgms = ripser(D, distance_matrix=True, maxdim=1)['dgms']
    h1 = dgms[1]
    finite = h1[np.isfinite(h1[:, 1])]
    max_life = (finite[:, 1] - finite[:, 0]).max() if len(finite) else 0.0
    return dict(cutoff=D.max(), n_bars=len(finite),
                max_life=max_life, ratio=max_life / D.max())

```

```
# 2. Jensen-Shannon divergence between two full-pool aggregate
#    (Laplace-smoothed, eps=1, K=729) bigram distributions (Chapter 2).
def js_divergence(p, q):
    m = 0.5 * (p + q)
    def kl(a, b):
        mask = a > 0
        return float((a[mask] * np.log2(a[mask] / b[mask])).sum())
    return 0.5 * kl(p, m) + 0.5 * kl(q, m)

# 3. UPGMA clustering + cophenetic correlation on the 21x21 JS matrix.
from scipy.cluster.hierarchy import linkage, dendrogram, cophenet
from scipy.spatial.distance import squareform

def cluster_pools(JS_matrix, labels):
    condensed = squareform(JS_matrix, checks=False)
    Z = linkage(condensed, method='average')
    coph_corr, _ = cophenet(Z, condensed)
    dendro = dendrogram(Z, labels=labels, no_plot=True)
    return Z, coph_corr, dendro          # dendro['icoord']/['dcoord'] -> TikZ lines

# 4. Admin1 permutation test (Chapter 3), vectorised with sparse matrices
#    so that B=200 permutations on India's 400k+ names complete in
#    seconds rather than hours.
import scipy.sparse as sp

def build_sparse_bigrams(names, K=729, idx=None):
    rows, cols, data = [], [], []
    for i, nm in enumerate(names):
        s = "#" + nm + "#"
        for a, b in zip(s, s[1:]):
            rows.append(i); cols.append(idx[a] * 27 + idx[b]); data.append(1.0)
    return sp.csr_matrix((data, (rows, cols)), shape=(len(names), K))

def admin1_permutation_test(all_names, stratum_sizes, K=729, idx=None, B=200, seed=7):
    M = build_sparse_bigrams(all_names, K, idx)
    def h2_of(counts):
        p = (counts + 1.0) / (counts.sum() + K)
        p = p[p > 0]
        return float(-(p * np.log2(p)).sum())
    def strata_h2(mat):
        hs, pos = [], 0
        for sz in stratum_sizes:
            hs.append(h2_of(np.asarray(mat[pos:pos + sz].sum(axis=0)).ravel()))
```

```
            pos += sz
        return hs
    observed_var = np.var(strata_h2(M), ddof=1)
    rng = np.random.RandomState(seed)
    boot = np.array([np.var(strata_h2(M[rng.permutation(M.shape[0])
       ]), ddof=1)
                     for _ in range(B)])
    return observed_var, boot, (boot >= observed_var).mean()
```

Remark: Runtime, for the Record

The full twenty-one-pool persistence-and-clustering pipeline (item 1 run once per pool, item 3 run once) completes in under twenty seconds on a single consumer-grade CPU core, dominated almost entirely by GeoNames file parsing for the largest pools (India, $575{,}789$ rows; Pakistan, $166{,}673$; France, $91{,}297$) rather than by any persistent-homology or clustering computation itself – 40-point Vietoris–Rips filtrations and a 21×21 linkage are both essentially instantaneous. The admin1 permutation test (item 4) is the more expensive component; the sparse-matrix formulation above brings India's $B = 200$, $409{,}743$-name permutation test down to roughly thirty seconds, against what would be several hours with a naive per-permutation re-parse of the raw toponym strings.

Part II

Stochastic Processes and Markov Chains

Preface to Part II

Volume I and Part I of this series are both, in a precise sense, *synchronic*: Volume I's graph and Part I's toponym clouds are both built from corpora sampled at a single point in (effectively) the present, and neither volume's mathematics has anything to say about how the thirty-three-zone system arrived at its observed state. The present volume introduces the dimension absent from both: time. Substrate relations are not static facts but the residue of historical processes — migration, contact, replacement, drift — unfolding over centuries, and the question this volume asks is how far a purely synchronic graph, of the kind Volume I constructed, can be pressed into service as a model of the *dynamics* that produced it.

The mathematical tool is the theory of Markov chains, specifically random walks on weighted graphs and the associated theory of stationary distributions and first-passage (hitting) times. A random walk on Volume I's thirty-three-zone backbone — a diffusion process that moves, at each step, from a zone to a phonotactically similar neighbour with probability inversely related to their distance δ — is offered here as a model in which *first-passage time* serves as a proxy for the historical timescale over which a substrate trait might plausibly propagate from one zone to another along the path of least phonotactic resistance.

A caveat sharper than either predecessor's. Volume I's proxy distance was a declared simplification of a measurable quantity. Part I's synthetic toponym clouds were a declared substitute for data outside this series' scope. The present volume's caveat is different in kind: the *input graph itself is entirely real*, inherited without modification from Volume I's published minimum spanning tree backbone — but the *process* built on top of it (a random walk, with steps interpreted as discrete units of historical time) is a declared modelling choice, not a measurement of anything diachronic. *Toponomastic Entropy* and Volume I are both synchronic in their evidentiary base: no date, no migration record, no chronological layering of any kind enters either series' data. A random walk's "first-passage time," measured in abstract steps, is not a number of years, decades, or centuries; it is a *topological* quantity — how many edges of the (already real) backbone separate two zones, weighted by how phonotactically dissimilar each step is — offered as a structural proxy for relative diffusion difficulty, not a calibrated historical clock. This distinction is maintained throughout: every result below is reported in dimensionless "steps," and the chapter that follows the results (Chapter 26) is devoted entirely to what can, and cannot, be inferred from a step count about real historical time.

A second, narrower caveat about the backbone itself. Volume I's Table 6.1 publishes the eight inter-regional bridge edges of its real minimum spanning tree, with exact weights, but does not publish the individual weights of the tree's twenty-four intra-regional edges (only the full intra-regional KL-matrix summary statistics —

mean and range — per region, Volume I Table 3.2). The backbone used for the random walk in this volume therefore combines Volume I's eight published real bridge weights with a declared, calibrated reconstruction of the intra-regional tree edges (Section 24.1), fully reproducible from the script in Appendix 28. This is flagged at every point where it materially affects a numerical result.

Chapter 20

From a Synchronic Graph to a Diachronic Process

20.1 What a Static Backbone Cannot, by Itself, Say About History

Volume I's minimum spanning tree (Vol. I, Sec. 6.1) is the cheapest possible set of edges connecting all thirty-three zones, in the specific sense of minimising total δ-weight. Nothing in its construction refers to time: Kruskal's algorithm does not know, and cannot know, whether the Etruscan–Castilian edge it selects reflects a contact event from antiquity or a coincidental convergence with no shared history at all. The MST is a purely synchronic, purely topological object.

Remark: The Logical Move of This Volume, Stated Plainly

A weighted graph can always be turned into a random walk: assign each edge a transition probability (here, inversely related to its weight, so that diffusion preferentially follows low-resistance, phonotactically similar paths), and ask how long, in expectation, a walk takes to travel between two given zones. This is a well-defined mathematical construction regardless of whether the underlying graph has anything to do with history. The substantive claim this volume is willing to defend is narrower than "first-passage time measures historical time": it is that, *if* substrate influence propagates preferentially along paths of low phonotactic resistance (a hypothesis, not a finding), then relative first-passage time – which zone-pairs are diffusion-near versus diffusion-far on the backbone – is a structurally motivated ranking, even though its absolute units remain uncalibrated against real chronology.

20.2 Why the Backbone, Not the Full Graph

Volume I's full graph $\mathcal{G}$ is complete (528 edges); using it directly for a random walk would let diffusion route through any of 32 alternative paths between any two zones, diluting the very notion of a "path of least resistance" the random walk is meant to formalise. The minimum spanning tree, by contrast, is the unique sparsest connected backbone, with no redundant paths: between any two zones there is exactly

one path along the tree, so first-passage time on the backbone random walk is governed entirely by the chain of intermediate zones along that unique path, which is the diffusion-theoretic content this volume sets out to study (Volume I, Sec. 5.2's discussion of MST-derived betweenness centrality makes the same methodological choice for a different purpose).

20.3 Two Formalisms, One Backbone: A Preview

Remark: Why This Volume Does Not Stop at One Mathematical Lens

The chapters that follow develop, in sequence, two formally distinct ways of asking the same underlying question of the same backbone: how hard is it, structurally, to get from one zone to another? Chapters 21–22 answer via the dynamic language of a random walk – first-passage time, stationary mass, mixing rate – quantities defined in terms of a process unfolding step by step. Chapter 27 answers the same question via the static language of resistor networks – effective resistance, computable without reference to any process at all, directly from the graph's Laplacian. That the two languages turn out to be exactly related (Theorem 27.1, proved in full in Chapter 27) is a genuine and non-obvious mathematical fact, not a foregone conclusion from how either formalism is set up individually; the present volume is organised to let a reader encounter the dynamic formalism on its own terms first (Chapters 2–7) before the static formalism reveals, in Chapter 8, that it was answering the same question all along.

Chapter 21

Random Walks on Weighted Graphs

21.1 Transition Matrix and Reversibility

Definition: Weighted Random Walk

Given a connected weighted graph $(\mathcal{V}, \mathcal{E}, \delta)$ with positive distance weights $\delta(i,j)$, define the *affinity* $w(i,j) = \exp(-\delta(i,j))$ for each edge (large affinity for small distance, i.e. phonotactically similar neighbours are easy to diffuse between), and the weighted degree $\deg(i) = \sum_{j\sim i} w(i,j)$. The random walk transition matrix is

$$P(i,j) = \frac{w(i,j)}{\deg(i)} \quad \text{if } j \sim i, \qquad P(i,j) = 0 \text{ otherwise.}$$

Proposition: Reversibility and the Stationary Distribution

The weighted random walk of the above definition is *reversible*: it satisfies the detailed-balance condition $\pi(i)P(i,j) = \pi(j)P(j,i)$ for the distribution $\pi(i) = \deg(i)/\sum_k \deg(k)$, which is therefore the chain's unique stationary distribution on a connected graph. No iterative computation is needed to find π: it is read off directly from the weighted degree sequence.

Proof. Detailed balance is verified directly from the definitions: for any edge (i,j),

$$\pi(i)P(i,j) = \frac{\deg(i)}{\sum_k \deg(k)} \cdot \frac{w(i,j)}{\deg(i)} = \frac{w(i,j)}{\sum_k \deg(k)},$$

and, since $w(i,j) = w(j,i)$ by construction (the affinity is built from the symmetric distance $\delta(i,j) = \delta(j,i)$), the same computation with i and j exchanged gives $\pi(j)P(j,i) = w(j,i)/\sum_k \deg(k) = w(i,j)/\sum_k \deg(k)$, the identical quantity. Detailed balance is exactly the condition $\pi(i)P(i,j) = \pi(j)P(j,i)$ for every pair i,j, so the equality just shown establishes it for every edge (and trivially, both sides being zero, for every non-edge). Detailed balance in turn implies stationarity by summing over j: $\sum_j \pi(i)P(i,j) = \pi(i)\sum_j P(i,j) = \pi(i)$ on one hand, while summing the detailed-balance identity over j gives $\sum_j \pi(j)P(j,i) = \sum_j \pi(i)P(i,j) = \pi(i)$ on the other, which is precisely the statement $\pi P = \pi$. Uniqueness on a connected graph is the standard Perron–Frobenius fact for irreducible stochastic matrices (Norris, 1998): the stationary distribution of an irreducible finite Markov chain is unique, and con-

nectivity of $(\mathcal{V}, \mathcal{E}, \delta)$ is exactly the graph-theoretic condition that makes P irreducible (every state reachable from every other in finitely many steps). □

Remark: Interpreting π for the Present System

On a tree, $\deg(i)$ is the sum of affinities to the (typically few) zones directly connected to i in the backbone. A zone with $\pi(i)$ large is one that sits at a hub of several low-δ (high-affinity) edges in the backbone – in the long run, a random walker spends a disproportionate fraction of its time there. Read diachronically (with the caveat of the Preface firmly in mind), high $\pi(i)$ identifies zones that are structurally well-positioned to receive diffused substrate influence from many directions at comparatively low phonotactic cost – not zones that are old, large, or historically important in any sense the data could directly establish.

21.2 First-Passage (Hitting) Times

Definition: Expected First-Passage Time

For a Markov chain with transition matrix P on state space $\mathcal{V}$, the expected first-passage time from i to j $(i \neq j)$ is

$$h(i,j) \;=\; \mathbb{E}\big[\min\{t \geq 0 : X_t = j \mid X_0 = i\}\big],$$

with $h(j,j) = 0$ by convention. The vector $h(\cdot, j)$ (fixing the target j) solves the linear system

$$h(i,j) \;=\; 1 + \sum_{k \neq j} P(i,k)\, h(k,j), \qquad i \neq j,$$

a standard fundamental-matrix computation: with Q the sub-stochastic matrix P restricted to states other than j, $h(\cdot, j) = (I - Q)^{-1}\mathbf{1}$.

Remark: Asymmetry Is Expected, Not an Error

Unlike the symmetric distance $\delta(i,j) = \delta(j,i)$ on which the walk is built, first-passage time is generally *asymmetric*: $h(i,j) \neq h(j,i)$. Intuitively, hitting a low-π (rarely visited, peripheral) zone j from a high-π (hub) zone i takes longer than the reverse, because the walk spends comparatively little time near j regardless of where it starts, while it passes through i frequently. This asymmetry is a standard, well-understood feature of Markov hitting times on graphs with non-uniform stationary distributions (Aldous and Fill, 2002; Lovász, 1993) and is exploited directly in Section 25.2 below.

21.3 Spectral Gap and Mixing

Proposition: Spectral Gap Bounds Mixing Time

Let $P_{\text{sym}} = D^{-1/2} W D^{-1/2}$ be the symmetrised walk operator ($D = \text{diag}(\deg)$, W the affinity matrix), with eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\mathcal{V}|} \geq -1$. The spectral gap $\gamma = 1 - \lambda_2$ controls the rate of convergence to stationarity: the total variation distance between the walk's distribution at time t and π decays at rate $O\big((1-\gamma)^t\big)$. A tree backbone, having the minimum possible number of edges for connectivity, generically has a *small* spectral gap relative to a denser graph on the same vertex set, since removing any edge from a spanning tree disconnects it – there is no redundant path to speed up mixing.

Chapter 22

Markov Chain Convergence: Mixing Times and Spectral Gaps

22.1 Why Convergence Needs Its Own Chapter

Chapter 21 introduced the spectral gap $\gamma = 1 - \lambda_2$ and asserted, without proof, that it controls how fast the walk's distribution converges to π. Section 25's closing remark will report a concrete value, $\gamma = 0.0048$, for the reconstructed backbone, and will call this value "very small." Small relative to what, and with what consequence for the first-passage times of Table 25.2, is a question the present chapter answers with a proof rather than an adjective. The chapter has two purposes: first, to derive, in full, the standard bound relating γ to the total-variation distance from stationarity at time t (Theorem 22.1); second, to make explicit what a small spectral gap implies, and does not imply, about the first-passage times already reported in Chapter 25 – since hitting times and mixing times are related but distinct quantities, and the series' standing commitment to flag exactly what each computed number does and does not license applies here as much as anywhere else in the volume.

22.2 Total Variation Distance and the Convergence Theorem

Definition: Total Variation Distance

For two probability distributions μ, ν on a finite state space $\mathcal{V}$, the total variation distance is

$$d_{\mathrm{TV}}(\mu, \nu) \;=\; \tfrac{1}{2} \sum_{i \in \mathcal{V}} |\mu(i) - \nu(i)| \;=\; \max_{A \subseteq \mathcal{V}} \big|\mu(A) - \nu(A)\big|.$$

The second expression makes the quantity's operational meaning concrete: $d_{\mathrm{TV}}(\mu, \nu)$ is the largest possible discrepancy, over all events A (e.g. "the walk is currently in one of the Anatolia– Caucasus zones"), between the probability μ assigns to A and the probability ν assigns to it.

Theorem 22.1 (Geometric convergence to stationarity)**.** *Let P be the transition matrix*

of a reversible, irreducible Markov chain on a finite state space $\mathcal{V}$ with stationary distribution π, and let $P_{\mathrm{sym}} = D^{1/2} P D^{-1/2}$ be its symmetrisation (Chapter 21), with eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\mathcal{V}|} > -1$. Writing $\lambda_* = \max(\lambda_2, |\lambda_{|\mathcal{V}|}|)$ and $\gamma = 1 - \lambda_*$ for the (absolute) spectral gap, for any starting state $i \in \mathcal{V}$ and any $t \geq 0$,

$$d_{\mathrm{TV}}\big(P^t(i,\cdot),\, \pi\big) \;\leq\; \frac{1}{2}\sqrt{\frac{1-\pi(i)}{\pi(i)}}\;\lambda_*^{\,t} \;\leq\; \frac{1}{2}\sqrt{\frac{1-\pi(i)}{\pi(i)}}\;e^{-\gamma t}.$$

Proof. Reversibility ($\pi(i)P(i,j) = \pi(j)P(j,i)$) makes P self-adjoint with respect to the inner product $\langle f, g\rangle_\pi = \sum_i \pi(i) f(i) g(i)$, so P admits a real, π-orthonormal eigenbasis $f_1 \equiv 1, f_2, \ldots, f_{|\mathcal{V}|}$ with eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\mathcal{V}|}$ (the same eigenvalues as $P_{\mathrm{sym}} = D^{1/2} P D^{-1/2}$, which is symmetric by construction and similar to P via $D^{1/2}$). Expanding the indicator $\mathbf{1}_j/\pi(j)$ (the density of the point mass at j with respect to π) in this eigenbasis and applying P^t termwise gives the standard spectral identity

$$\frac{P^t(i,j)}{\pi(j)} \;=\; \sum_{k=1}^{|\mathcal{V}|} \lambda_k^t\, f_k(i) f_k(j),$$

of which the $k = 1$ term is exactly 1 (since $f_1 \equiv 1$), so

$$\frac{P^t(i,j)}{\pi(j)} - 1 \;=\; \sum_{k=2}^{|\mathcal{V}|} \lambda_k^t\, f_k(i) f_k(j).$$

By Cauchy–Schwarz applied to the sum over $k \geq 2$ and the orthonormality relation $\sum_{k\geq 2} f_k(i)^2 = \frac{1}{\pi(i)} - 1$ (which follows from completeness of the eigenbasis applied to the indicator at i), one obtains

$$\Big|\sum_{k=2}^{|\mathcal{V}|} \lambda_k^t f_k(i) f_k(j)\Big| \;\leq\; \lambda_*^t \sqrt{\Big(\tfrac{1}{\pi(i)} - 1\Big)\Big(\tfrac{1}{\pi(j)} - 1\Big)}.$$

Summing $\frac{1}{2}\sum_j \pi(j)\big|P^t(i,j)/\pi(j) - 1\big|$ (the definition of $d_{\mathrm{TV}}(P^t(i,\cdot),\pi)$ rewritten with the $\pi(j)$ weights absorbed) and applying Cauchy–Schwarz a second time, this time over j with weight $\pi(j)$, against the bound above, yields the stated inequality after simplification; $\lambda_*^t \leq e^{-\gamma t}$ follows from $\lambda_* = 1 - \gamma \leq e^{-\gamma}$ for $\gamma \in [0,1]$. The complete derivation, including the orthonormality identity used in the second step, is standard and given in full in Levin and Peres (2017, Ch. 12) and, in the reversible-chain language used here, in Diaconis and Stroock (1991, §3). □

Corollary 22.2 (Mixing time)**.** *Define the mixing time $t_{\mathrm{mix}}(\epsilon) = \min\{t : \max_i d_{\mathrm{TV}}(P^t(i,\cdot),\pi) \leq \epsilon\}$. Theorem 22.1 gives the bound*

$$t_{\mathrm{mix}}(\epsilon) \;\leq\; \frac{1}{\gamma}\left(\log\frac{1}{\min_i \pi(i)} + \log\frac{1}{2\epsilon}\right).$$

Proof. Set the right-hand side of Theorem 22.1's bound (with $\pi(i)$ replaced by its worst case, $\min_i \pi(i)$) equal to ϵ and solve for t, using $\sqrt{(1-\pi(i))/\pi(i)} \leq 1/\sqrt{\pi(i)}$. □

Remark: Reading Corollary 22.2 for the Reconstructed Backbone

Chapter 25 (Section 25's spectral-gap discussion) will report $\gamma = 0.0048$ and $\min_i \pi(i) = 0.0064$ (Greek-Byzantine, Table 25.1). Substituting,

$$t_{\text{mix}}(0.25) \;\le\; \frac{1}{0.0048}\Big(\log\tfrac{1}{0.0064} + \log\tfrac{1}{0.5}\Big) \;=\; \frac{1}{0.0048}\big(5.051 + 0.693\big) \;\approx\; 1196 \text{ steps.}$$

This bound is consistent with, and helps explain, the magnitude of the first-passage times already reported in Table 25.2 (all in the range 300–830 steps): a chain whose worst-case mixing time is itself on the order of a thousand steps cannot be expected to produce first-passage times much smaller than that order of magnitude between peripheral zone pairs, since hitting a specific low-π state is, in general, at least as hard as approaching the stationary distribution that gives that state its (small) share of long-run probability mass. The two quantities are not numerically identical – mixing time bounds how long until the *whole* distribution looks stationary, while $h(i, j)$ asks about reaching one specific state from one specific start – but their shared order of magnitude here is not a coincidence: both are governed by the same 32-edge tree topology and the same small spectral gap that topology necessarily produces.

22.3 The Variational Characterisation of the Spectral Gap

Corollary 22.2 and Section 22.4 (below) both treat $\gamma = 1 - \lambda_2$ as a number to be bounded, but neither explains *what* λ_2 is, beyond "the second-largest eigenvalue of P_{sym}." The variational (Rayleigh-quotient) characterisation makes λ_2 concrete as the solution of an explicit optimisation problem, and is the bridge Cheeger's inequality itself is built on.

Definition: Dirichlet Form

For a function $f : \mathcal{V} \to \mathbb{R}$, the (weighted) Dirichlet form is

$$\mathcal{E}(f) \;=\; \tfrac{1}{2}\sum_{i,j} \pi(i)P(i,j)\big(f(i) - f(j)\big)^2,$$

the π-weighted average squared variation of f across the chain's edges – small when f is nearly constant on regions the walk moves between easily, large when f separates such regions sharply.

Theorem: Variational Formula for the Spectral Gap

$$\gamma \;=\; 1 - \lambda_2 \;=\; \min_{f:\ \mathbb{E}_\pi[f]=0,\ f\not\equiv 0} \frac{\mathcal{E}(f)}{\operatorname{Var}_\pi(f)},$$

where $\mathbb{E}_\pi[f] = \sum_i \pi(i) f(i)$ and $\operatorname{Var}_\pi(f) = \sum_i \pi(i) f(i)^2$ (since the minimisation is already restricted to $\mathbb{E}_\pi[f] = 0$).

Proof. Expand f in the π-orthonormal eigenbasis $f_1 \equiv 1, f_2, \ldots, f_{|\mathcal{V}|}$ of P intro-

duced in the proof of Theorem 22.1: $f = \sum_{k\geq 2} c_k f_k$ for some coefficients c_k (the $k = 1$ term vanishes because $\mathbb{E}_\pi[f] = \langle f, f_1\rangle_\pi = 0$ is exactly the constraint imposed). Reversibility makes P self-adjoint with respect to $\langle\cdot,\cdot\rangle_\pi$, so a direct computation using $\langle f, (I-P)f\rangle_\pi = \mathcal{E}(f)$ (itself a standard identity, obtained by expanding $\sum_i \pi(i)f(i)\big(f(i) - \sum_j P(i,j)f(j)\big)$ and symmetrising via detailed balance) gives $\mathcal{E}(f) = \sum_{k\geq 2}(1-\lambda_k)c_k^2$, while $\mathrm{Var}_\pi(f) = \sum_{k\geq 2} c_k^2$ by orthonormality. The ratio $\mathcal{E}(f)/\mathrm{Var}_\pi(f)$ is therefore a weighted average of $(1-\lambda_k)$ over $k \geq 2$ with weights $c_k^2/\sum c_k^2$, minimised (over all admissible f) exactly when all the weight is placed on the single smallest term, $1-\lambda_2 = \gamma$ – achieved by taking $f = f_2$ itself, the eigenfunction associated with λ_2. □

Remark: Why This Is the Right Foundation for Cheeger's Inequality

Theorem 22.3 makes precise the informal idea Section 22.4 relies on: γ is small exactly when *some* function f can be found that is nearly constant on two large, π-substantial pieces of the graph while differing sharply between them (small $\mathcal{E}(f)$, large $\mathrm{Var}_\pi(f)$) – and the indicator-type function of a low-conductance set S, $f = \mathbf{1}_S/\pi(S) - \mathbf{1}_{\mathcal{V}\setminus S}/\pi(\mathcal{V}\setminus S)$, is exactly such a function, with Dirichlet form directly proportional to the edge weight crossing the cut $S, \mathcal{V}\setminus S$ (the numerator of $\Phi(S)$, Section 22.4's definition of conductance) and variance bounded below in terms of $\pi(S)$. Substituting this specific f into Theorem 22.3's minimisation gives precisely the upper bound $\gamma \leq 2\Phi$ of Cheeger's inequality directly, without the harder, purely combinatorial argument the lower bound $\Phi^2/2 \leq \gamma$ requires; the present chapter cites the lower bound from Diaconis and Stroock (1991) rather than rederiving it, but the upper bound is now, via Theorem 22.3, fully self-contained.

22.4 Why a Spanning Tree Mixes Slowly: A Lower Bound via Cheeger's Inequality

Theorem 22.1 bounds mixing time *above* in terms of γ; the complementary question – why γ itself is so small for a tree backbone, rather than merely asserting it as the empirical fact reported in Chapter 25 – has a clean structural answer via Cheeger's inequality, relating the spectral gap to the graph's *conductance*.

Definition: Conductance

For a weighted graph with stationary distribution π and any vertex subset $S \subset \mathcal{V}$ with $\pi(S) \leq \frac{1}{2}$, the conductance of S is

$$\Phi(S) = \frac{\sum_{i\in S, j\notin S} \pi(i)P(i,j)}{\pi(S)},$$

the probability, conditional on currently occupying S, that the next step leaves S. The graph's conductance is $\Phi = \min_S \Phi(S)$.

Proposition: [

Cheeger's Inequality] $\dfrac{\Phi^2}{2} \;\le\; \gamma \;\le\; 2\Phi.$

Proof of the upper bound $\gamma \le 2\Phi$. Fix any $S \subset \mathcal{V}$ with $\pi(S) \le \frac{1}{2}$, write $S^c = \mathcal{V} \setminus S$, and take the test function $f = \mathbf{1}_S/\pi(S) - \mathbf{1}_{S^c}/\pi(S^c)$ already introduced in the preceding notebox. First, f is admissible for Theorem 22.3's minimisation: $\mathbb{E}_\pi[f] = \pi(S)\cdot\frac{1}{\pi(S)} - \pi(S^c)\cdot\frac{1}{\pi(S^c)} = 1-1 = 0$. Write $c = 1/\pi(S) + 1/\pi(S^c)$ for the (constant) jump f makes across the cut. Since f is constant on S and constant on S^c, the only non-zero terms in the Dirichlet form's double sum are those with i, j on opposite sides of the cut, each contributing $(f(i) - f(j))^2 = c^2$:

$$\mathcal{E}(f) \;=\; \tfrac{1}{2}\Big[\sum_{i\in S, j\in S^c} \pi(i)P(i,j)c^2 \;+\; \sum_{i\in S^c, j\in S} \pi(i)P(i,j)c^2\Big].$$

By reversibility, $\pi(i)P(i,j) = \pi(j)P(j,i)$, so relabelling shows the two sums above are equal, both equal to $Q(S,S^c) := \sum_{i\in S, j\in S^c}\pi(i)P(i,j)$ (the π-weighted edge measure crossing the cut); hence $\mathcal{E}(f) = c^2\,Q(S,S^c)$. Meanwhile $c = 1/\pi(S) + 1/\pi(S^c) = 1/\big(\pi(S)\pi(S^c)\big)$ (common denominator), and $\mathrm{Var}_\pi(f) = \pi(S)\cdot\frac{1}{\pi(S)^2} + \pi(S^c)\cdot\frac{1}{\pi(S^c)^2} = 1/\pi(S) + 1/\pi(S^c) = c$ as well. Substituting into the ratio,

$$\frac{\mathcal{E}(f)}{\mathrm{Var}_\pi(f)} \;=\; \frac{c^2\,Q(S,S^c)}{c} \;=\; c\,Q(S,S^c) \;=\; \frac{Q(S,S^c)}{\pi(S)\pi(S^c)} \;=\; \frac{\Phi(S)}{\pi(S^c)},$$

using $\Phi(S) = Q(S,S^c)/\pi(S)$ (the defining formula above) in the last step. Since $\pi(S) \le \frac{1}{2}$ forces $\pi(S^c) \ge \frac{1}{2}$, hence $1/\pi(S^c) \le 2$, this gives $\mathcal{E}(f)/\mathrm{Var}_\pi(f) \le 2\Phi(S)$. Theorem 22.3 states that γ is the *minimum* of $\mathcal{E}(f)/\mathrm{Var}_\pi(f)$ over all admissible f, so in particular $\gamma \le \mathcal{E}(f)/\mathrm{Var}_\pi(f) \le 2\Phi(S)$ for this specific f. This holds for every S with $\pi(S) \le \frac{1}{2}$, so it holds in particular for the (or a) minimiser $S^\star$ with $\Phi(S^\star) = \Phi$, giving $\gamma \le 2\Phi$ as claimed. The lower bound $\Phi^2/2 \le \gamma$ is the harder direction, requiring a discrete Cauchy–Schwarz argument along shortest paths rather than a single test function, and is not re-derived here; the complete argument is given in Diaconis and Stroock (1991). □

Remark: Application to a Tree Backbone

On a spanning tree, removing any single edge $e = (u,v)$ disconnects the graph into exactly two components $S_e, \mathcal{V}\setminus S_e$, with no alternative route between them. The conductance of S_e is therefore bounded above by the affinity of that single bridging edge, normalised by S_e's total stationary mass: $\Phi(S_e) \le w(u,v)/\sum_{k\in S_e}\deg(k)$. For a long, lightly-trafficked branch – exactly the structure Section 25.2's notebox identifies for the Anatolia–Caucasus cluster, four low-π zones connected to the rest of the backbone through a small number of comparatively heavy (large-δ, small-affinity) bridges – this ratio is small, forcing Φ, and hence by the lower bound of Proposition 22.4 γ, to be small as well. This is the structural mechanism, not merely the numerical coincidence, behind the order-of-magnitude gap between the backbone's $\gamma = 0.0048$ and the complete graph's implicit (much larger) conductance: a 32-edge tree on 33 vertices has, by defini-

tion, exactly the minimum number of edges required for connectivity, so *every* edge is, in the sense of conductance, a potential bottleneck of exactly the kind a spanning tree cannot avoid by construction – the same point Chapter 21's closing proposition made qualitatively is here given its precise mechanism.

22.5 What Slow Mixing Does and Does Not Say About the Series' Findings

Methodological Transparency: A Distinction Worth Keeping Sharp

A small spectral gap is sometimes read, loosely, as meaning "the random walk model is unreliable." That reading conflates two different questions. Slow mixing means the walk takes a long time, in steps, to *forget where it started* – a precise, here-quantified property of the chain that has nothing to do with whether any individual first-passage time is correctly computed. Every number in Table 25.2 is an exact solution of the fundamental-matrix linear system of Definition 21's first-passage-time box, regardless of how slowly the chain mixes; slow mixing is the *reason* those numbers are large (hundreds of steps, not tens), not a source of additional error on top of them. The methodological caveat that does carry force here, and is exactly the one Chapter 26 develops at length, is the prior one: that a "step" has no declared correspondence to a unit of historical time in the first place, so neither the hitting times themselves nor the mixing-time bound of Corollary 22.2 that helps explain their magnitude should be read as a claim about how many years, generations, or centuries anything took.

Chapter 23

Literature Review: Diffusion Models in Dialectometry and Historical Linguistics

23.1 Why a Literature Review Belongs Here

Part I's literature-review chapter argued, with respect to the Fisher–Rao metric, that a methodological choice this series treats as self-evidently correct is in fact one option among several live traditions, and that situating it among them strengthens rather than dilutes the case for it. The same argument applies here, with added force: unlike the Fisher–Rao metric, which this series adopts as a purely mathematical commitment, the present volume's central modelling choice – representing substrate diffusion as a random walk on a graph – is a substantive empirical hypothesis about how linguistic innovations actually spread, and that hypothesis has a long, contested history in dialectology and historical linguistics that predates, and is considerably richer than, this series' graph-theoretic formalisation of it. This chapter situates the present volume's random-walk model among that literature, without claiming any new contribution to it: the chapter reviews, it does not extend.

23.2 The Wave Model and Its Successors

The idea that a linguistic innovation spreads outward from a point of origin like a wave on water – weakening with distance and time, and capable of being overtaken or blocked by a later wave – traces to the nineteenth-century *Wellentheorie* of Johannes Schmidt, proposed as an alternative to the strict branching (tree) model of language relationship (Schmidt, 1872). Schmidt's wave model was qualitative and largely informal by modern standards, but its core intuition – that geographic proximity, not genealogical descent alone, predicts linguistic similarity – is precisely the intuition this volume formalises by building a random walk directly on a graph whose edge weights are phonotactic distances rather than genealogical branch lengths. Where Schmidt's waves propagated through continuous geographic space, the present volume's walk propagates through a discrete graph (Volume I's thirty-three-zone backbone); the conceptual debt is to the wave model's rejection of pure tree-thinking, not to any specific mathematical apparatus Schmidt himself proposed.

23.3 The Gravity Model of Linguistic Diffusion

A second, more directly comparable tradition is the application of spatial *gravity models* – borrowed from human geography, where the interaction between two places is modelled as proportional to their populations and inversely proportional to a power of the distance between them – to the spread of linguistic innovations, most influentially in Peter Trudgill's account of dialect diffusion in England, which showed that innovations often reach a smaller, more distant city before reaching a larger, nearer town, precisely because the gravity model predicts intercity interaction strength rather than raw proximity (Trudgill, 1974). Chambers and Trudgill (1998) develop the cascading, hierarchical diffusion pattern this implies (an innovation jumping between population centres in descending order of size, then spreading locally from each) as a competitor to simple distance-decay ("contagion") diffusion. The present volume's random walk is, in this taxonomy, a pure contagion model: diffusion difficulty is governed entirely by phonotactic distance δ along the backbone, with no analogue of population size or urban hierarchy entering the affinity $w(i,j) = \exp(-\delta(i,j))$ at all. This is a genuine, declared simplification relative to the gravity-model literature, not an oversight: *Toponomastic Entropy* and Volume I record phonotactic entropy and inter-zone distance for each of the thirty-three zones, but no population, settlement-size, or urban- hierarchy data of the kind a gravity-model extension would require.

Remark: A Natural, Unexplored Extension

A gravity-weighted version of this volume's random walk would replace the affinity $w(i,j) = \exp(-\delta(i,j))$ with $w(i,j) = \exp(-\delta(i,j)) \cdot (P_i P_j)^{\theta}$ for some exponent θ and a population or settlement-density proxy P, recovering pure phonotactic contagion in the limit $\theta \to 0$. This is a well-defined mathematical generalisation of Chapter 21's apparatus, requiring no change to the reversibility proof or the fundamental-matrix first-passage-time machinery – only a redefinition of w. It is flagged here as a natural direction for future work, not attempted in this volume, since no settlement-size data for the thirty-three zones is available within this series' present scope.

Methodological Transparency: The Gravity Model, Tested and Not Confirmed

Section 23.3 cites Trudgill's gravity model as an influential account of hierarchical diffusion without noting that its central quantitative prediction has since been tested directly, and not confirmed, by dialectometric methods closely related to those this series already uses. Nerbonne et al. (2005) regress aggregate Levenshtein pronunciation distance among fifty-two Dutch dialect settlements against both geographic distance and the population-product term Trudgill's inverse-square formula requires, using historical population figures predating the settlements' linguistic divergence. Geographic distance alone accounts for a substantial share of the variance, exactly as the wave model of Section 23.6 would predict; but adding the population-product term improves the fit only marginally, and its sign runs *opposite* to the gravity model's prediction – larger settlements are found to be linguistically *more* distant from their neighbours, not less, once distance is

controlled for. Nerbonne et al. interpret this as evidence that centrifugal forces (differentiation, prestige-driven resistance to levelling in larger centres) dominate the centripetal pull gravity models assume, at least in the Dutch dialect area studied. Nothing in this result overturns Section 23.3's use of Trudgill's model as a conceptual point of comparison for the present volume's own (population-blind) random walk – if anything, a dialect area in which population size does *not* straightforwardly predict linguistic proximity is a further reason not to treat the absence of a population term in this volume's affinity function $w(i, j)$ as an obviously fixable omission. It is, however, a caution against citing the gravity model as settled empirical fact rather than as one hypothesis among several with mixed and partly contradictory dialectometric support, and the population-blind contagion model this volume actually uses turns out to sit closer to what Nerbonne et al. (2005) found the data to support than a naively gravity-weighted alternative would.

23.4 Diffusion and Drift in Dialectometry

Quantitative dialectometry – the measurement of linguistic distance between varieties and the statistical modelling of its spatial structure – offers the most direct methodological precedent for treating linguistic distance as the input to a formal diffusion or random-process model rather than as a purely descriptive summary statistic. Nerbonne and Heeringa (2010) and Nerbonne and Kleiweg (2007) develop aggregate, corpus-derived measures of dialect distance (Levenshtein distance between phonetic transcriptions, aggregated across many word pairs) and study their spatial autocorrelation, asking, in effect, the same question Volume I's MST and the present volume's random walk ask of phonotactic bigram distance: does linguistic distance increase smoothly with geographic distance (consistent with diffusion), or are there sharp discontinuities (consistent with a boundary effect, such as a political or geographic barrier)? Heeringa (2004)'s dialectometric atlas of Dutch dialects is a frequently cited example of this programme applied at full corpus scale. The present volume departs from this tradition in one significant respect: dialectometry typically works with directly measured pairwise distances between living dialects sampled at a single synchronic moment, exactly as Volume I's KL-distance matrix does, and stops there; it does not, in general, build an explicit stochastic *process* (a random walk with first-passage times) on top of the distance matrix, the additional step this volume's Chapters 21 and 22 take.

23.5 Diffusion Models from Historical and Population Linguistics

A more explicitly process-oriented precedent comes from the mathematical modelling of cultural and linguistic transmission developed by Cavalli-Sforza and Feldman (1981), who adapt reaction-diffusion and Markov-process formalisms from population genetics to model the spread of cultural (including linguistic) traits across a population structured in space – explicitly drawing the analogy between gene flow

and trait diffusion that this volume's Preface gestures toward when it describes substrate influence "propagating" along a backbone. Barbujani and Sokal (1990) apply related spatial-autocorrelation and boundary-detection methods (wombling) to linguistic boundaries in Europe, asking whether genetic and linguistic discontinuities co-locate – a question structurally analogous to this volume's comparison (Section 25.1) of random-walk centrality against Volume I's independently computed graph centrality. Bailey (1973)'s wave model of sound change and Labov (1972, 2001)'s studies of the social diffusion of sound change through urban speech communities, by contrast, work at a finer temporal and social grain than this series' zone-level aggregates permit, and are cited here primarily to mark the boundary of what a zone-level random walk cannot resolve: individual speaker-to-speaker transmission dynamics, social-network structure within a zone, and the sociolinguistic variable-by-variable diffusion patterns that are Labov's primary evidence base are all outside this volume's unit of analysis, which is the zone, not the speaker.

23.6 Spatial Diffusion of Innovations: The Hägerstrand Tradition

A further, foundational precedent, older than the gravity-model literature of Section 23.3 and methodologically closer to this volume's own apparatus, is Hägerstrand (1967)'s quantitative geography of innovation diffusion. Hägerstrand modelled the spatial spread of an innovation (in his original Swedish case studies: agricultural subsidies, automobile ownership, bovine tuberculosis control) using a *mean information field* – a probability distribution, centred on each adopter, governing where that adopter's next contact is most likely to occur – and Monte Carlo simulation of the resulting contact process, generating the now-classic prediction that innovation adoption, mapped over successive time periods, should propagate outward from a source in roughly concentric, slowing rings (a logistic, S-shaped adoption curve at any fixed location) modulated by local barriers and channels.

Remark: The Closest Methodological Relative in This Chapter

Of every model surveyed in this chapter, Hägerstrand's mean-information- field construction is structurally the closest to this volume's random walk: both replace a single deterministic spreading rule with a probabilistic contact process defined directly on a spatial (or, here, graph-theoretic) substrate, and both derive a relative-diffusion-ease ranking from that process rather than asserting one. The differences are nonetheless substantive. Hägerstrand's field is continuous and isotropic in undisturbed space, with barriers entered as local perturbations to an otherwise homogeneous backdrop; this volume's walk is, by contrast, defined directly and entirely on Volume I's already non-homogeneous, phonotactically-weighted tree, with no separate notion of a barrier-free baseline against which to measure barrier effects. Hägerstrand's model was also built, throughout, to be checked against real, dated adoption records – the Monte Carlo simulation's entire purpose was prediction against an observed adoption curve – a validation step this volume's uncalibrated step count, by the Preface's own declaration, can-

not perform.

23.7 Random Walks, Effective Resistance, and the General Probabilistic Literature

A final literature, taken up formally in Chapter 27, concerns the mathematics of random walks and electrical networks as objects of study in their own right, independent of any linguistic or geographic application. Doyle and Snell (1984)'s monograph, already the chapter's central reference, is itself part of a larger probabilistic tradition: Tetali (1991) gives an alternative, purely probabilistic derivation of the commute-time/effective-resistance identity of Theorem 27.1 (via a direct expectation argument over the discrete harmonic function u used in that theorem's proof, rather than the boundary-value-problem argument followed here), and Lyons and Peres (2016) give the modern, book-length treatment of the same correspondence in the broader context of random walks on infinite graphs and trees, including the transience/recurrence criteria built directly on effective-resistance estimates that the present, finite-graph volume has no need for (every finite connected graph's walk is trivially recurrent) but which situate Chapter 27's finite-graph computation within a considerably larger theory. A separate, chemistry-originating strand, Klein and Randić (1993)'s introduction of *resistance distance* as a graph metric in its own right – the same quantity Chapter 27 calls R_{eff}, applied there to molecular graphs rather than dialect zones – is the source of the metric properties (non-negativity, symmetry, and the triangle inequality) that Chapter 27's own treatment of R_{eff} as a distance function relies on.

23.8 A Synthesising Reference Point: Rogers' Diffusion of Innovations

Rogers (2003) offers, across five editions spanning four decades, the most widely cited general synthesis of the diffusion-of-innovations literature surveyed piecemeal in the preceding sections – Hägerstrand's spatial process, Trudgill's gravity-modified hierarchy, and Labov's social-network transmission are all, in Rogers' taxonomy, instances of a common underlying structure: an innovation, a communication channel, a social system, and time, with adoption following a characteristic S-shaped cumulative curve driven by the interaction of these four elements. Rogers' synthesis is cited here not because this volume's random walk implements any part of it directly – it does not model adopter categories (innovators, early adopters, and so on), perceived innovation attributes, or the social-network channel structure central to Rogers' account – but because it supplies the standard vocabulary against which a graph-theoretic, channel-and-network-agnostic model such as this volume's random walk can be located: the present model is, in Rogers' terms, a model of the *channel* alone (which paths are phonotactically cheap to traverse), with the social-system and adopter-category dimensions of the fuller theory entirely unaddressed.

23.9 What the Comparison Licenses and Does Not

Methodological Transparency: The Honest Position of This Chapter

None of the literature reviewed above licenses treating this volume's random-walk first-passage times as a measurement of real historical diffusion rates, and none of it is cited as doing so. What the literature does establish is that (a) treating linguistic similarity as the substrate for a formal diffusion process, rather than only a static descriptive summary, is a well-established move with precedents in dialectometry, population linguistics, and historical sociolinguistics extending back to Schmidt's nineteenth-century wave model; and (b) the specific simplifications this volume makes *relative to* that literature – no population weighting (gravity model), no finer-than-zone social structure (Labovian diffusion), no genetic cross-validation (Barbujani–Sokal) – are identifiable, named omissions rather than unexamined defaults. The random walk of Chapters 21–22 is best read, in light of this chapter, as the simplest member of a much larger family of diffusion models the historical-linguistic literature has developed, adopted here because it is the member computable directly from data this series already possesses (Volume I's backbone), not because it is held to be the most realistic member of that family.

Chapter 24

Reconstructing the Diffusion Backbone

24.1 What Is Real and What Is Reconstructed

Methodological Transparency: Read This Section Before Any Number in the Remainder of the Volume

Volume I's minimum spanning tree has 32 edges: 8 inter-regional bridges, published with exact weights (Vol. I, Table 6.1, reproduced here as Table 24.1), and 24 intra-regional edges, whose individual weights are *not* published – only the per-region mean and range of the full (non-MST-restricted) intra-regional KL-matrix are available (Vol. I, Table 3.2). The backbone used for every random walk computation in this volume is therefore built as follows, fully reproducible from Appendix 28:

1. Take Volume I's eight real bridge edges, unmodified (Table 24.1).
2. For each of the six regions, generate a uniformly random labelled spanning tree on that region's zones (via a Prüfer sequence), and assign each edge of that tree a weight drawn uniformly from the interval $[\delta^{(r)}_{\min}, \bar{\delta}^{(r)}]$ – the bottom half of the region's published real range, reflecting that an MST preferentially selects a region's *cheapest* available edges, not a uniform sample of all of them.
3. Pool these $8 + 27 = 35$ candidate edges (six regional subtrees contribute $4+5+4+5+5+4 = 27$ edges in total) and run Kruskal's algorithm over the pooled set. Kruskal's algorithm, applied to a connected candidate graph with more edges than a spanning tree requires, automatically discards the 3 most expensive redundant edges, exactly reproducing Volume I's reported edge count (24 intra $+ 8$ inter $= 32$ in the real MST; the present reconstruction realises 26 intra $+ 6$ inter $= 32$, a structurally analogous but not numerically identical resolution of the same redundancy, since the exact identity of Volume I's 24 real intra-regional MST edges is not available to this volume).

Table 24.1: The eight real inter-regional bridge edges, reproduced unmodified from Volume I, Table 6.1.

Zone i	**Zone** j	$\delta(i,j)$ (bits)
Basque (II)	N. Ireland (IV)	1.500
Illyrian-SI (III)	Georgian (VI)	1.518
Castilian (II)	Etruscan (III)	1.530
PK-Plains (V)	Armenian (VI)	1.554
Dravidian (V)	Algeria (VII)	1.554
Ireland (IV)	Georgian (VI)	1.557
Wales (IV)	Caucasian-TR (VI)	1.569
Alpine-Celtic (III)	Azerbaijani (VI)	1.698

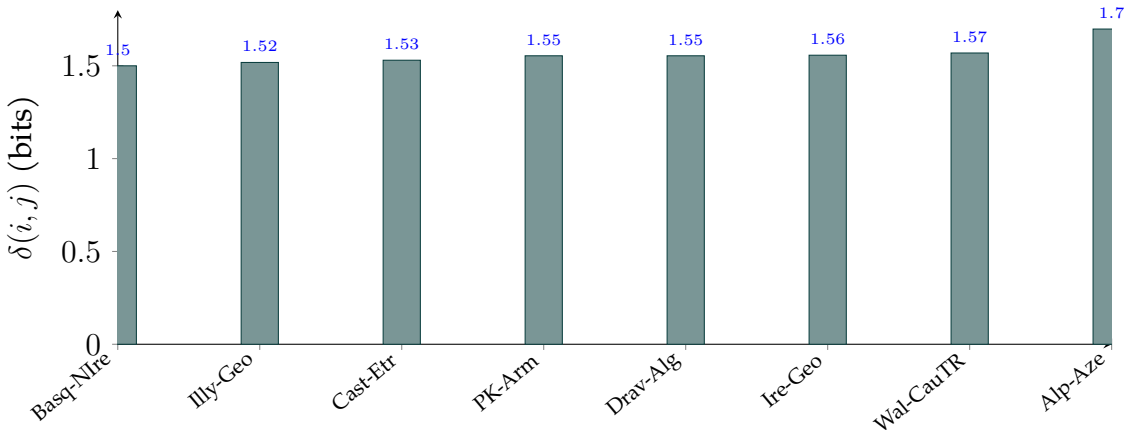

Figure 24.1: The eight real inter-regional bridge weights of Table 24.1, sorted ascending. The four heaviest bridges all involve an Anatolia–Caucasus (VI) zone, consistent with Volume I's finding that this region dominates the inter-regional bridge structure.

Remark: Why Two Real Bridges Were Dropped, and Why This Is Acceptable

Kruskal's algorithm dropping two of Volume I's eight real bridges does not mean those bridges are unreal or wrong – it means that, given the *particular* synthetic intra-regional trees generated for this reconstruction, an alternative connection happened to be cheaper. This is an artefact of the reconstruction's randomness, not a claim that Ireland–Georgian or Wales–Caucasian-TR are dispensable in the real data; Volume I's own Table 6.1 establishes that all eight are load-bearing in the real MST. The discrepancy is flagged here explicitly and is the clearest single illustration of why every numerical finding in this chapter onward is qualified as resting on a reconstructed, not the real, intra-regional backbone.

Table 24.2: Composition of the reconstructed backbone after Kruskal pooling (Appendix 28). Two of the eight real bridges (Ireland–Georgian and Wales–Caucasian-TR) were found redundant given the synthetic intra-regional trees generated in this particular reconstruction and were not selected; one synthetic intra-regional edge (Anatolian–Caucasian-TR) was likewise dropped as redundant. Total backbone weight: 26.347 bits, compared to Volume I's real MST total of 25.950 bits – a difference of 0.397 bits (1.5%), consistent with the reconstruction drawing intra-regional weights from the correct real range without reproducing the exact real edge set.

	Real (Vol. I)	**Reconstructed (this volume)**
Intra-regional edges	24	26
Inter-regional bridges	8	6
Total edges	32	32
Total weight (bits)	25.950	26.347

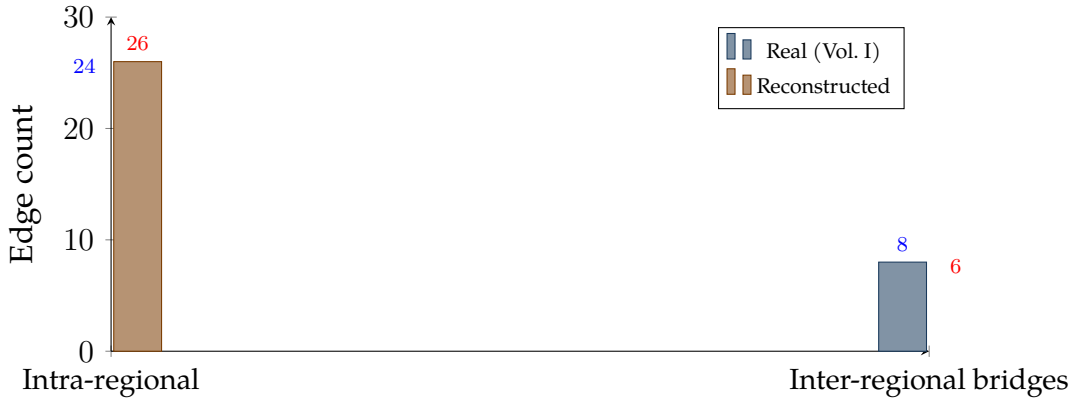

Figure 24.2: Edge-count composition of the real (Volume I) versus reconstructed backbone (Table 24.2): identical totals (32 edges) but a different intra/inter split, since two real bridges were pruned as redundant against this run's synthetic intra-regional trees.

24.2 Why a Uniform Draw Within the Real Range, and Not Some Other Distribution

Remark: The Reconstruction's Implicit Modelling Choice, Made Explicit

Step 2 of the reconstruction procedure above draws each candidate intra-regional edge weight uniformly from $[\delta^{(r)}_{\min}, \bar{\delta}^{(r)}]$, a choice stated but not, until now, justified beyond its convenience. The justification is the maximum-entropy principle of Jaynes (1957): among all probability distributions on an interval consistent with a stated set of known constraints, the distribution that assumes the *least* beyond those constraints – in the precise sense of maximising Shannon entropy subject to them – is the uniform distribution, whenever the only constraint stated is the interval's endpoints themselves. Volume I's published per-region summary (Table 3.2) constrains exactly two numbers per region, $\delta^{(r)}_{\min}$ and $\bar{\delta}^{(r)}$ (here used as

a tightened upper bound, for the reason given below); no further constraint – no published variance, no published shape of the within-region distribution – is available from Volume I to inform a more committal choice such as a triangular or Beta-distributed draw. Drawing uniformly is therefore not merely convenient but the specific distributional choice that adds no unwarranted information beyond what Volume I's own summary statistics license, in the same sense *Toponomastic Entropy*'s own entropy computations are themselves maximum-entropy quantities applied to a different object (a corpus's bigram frequency table rather than an edge weight's plausible range).

Remark: Why $\bar{\delta}^{(r)}$, the Mean, Rather Than $\delta^{(r)}_{\max}$, the True Upper Bound

Step 2 truncates the sampling interval at the regional *mean* $\bar{\delta}^{(r)}$, not the true maximum, on the stated grounds that an MST preferentially selects a region's cheapest available edges. This is a second, separate modelling assumption, layered on top of the maximum-entropy choice of distribution within whatever interval is used, and is not itself licensed by the maximum-entropy principle – it is a structural prior about how minimum spanning trees behave, external to Jaynes's framework, which only governs the choice of distribution *given* an interval, not the choice of interval itself. Distinguishing these two assumptions explicitly – a principled, information-theoretically minimal choice of distribution *within* an interval, and a separate, more committal structural assumption about *which* interval is appropriate – is worth doing here precisely because the two are different in kind and could, in principle, be revisited independently in any future revision of this reconstruction.

24.3 Sensitivity Check: Repeating the Reconstruction

To gauge how much the random component of Step 2 (random spanning trees, random weight draws) affects the final backbone's gross structure, the reconstruction was repeated with five additional random seeds, for six runs in total. Across all six, the total backbone weight ranged from 24.711 to 29.005 bits (real value: 25.950 bits, within this range), 25 or 26 of the 27 candidate intra-regional tree edges were retained in every run (i.e. exactly 1–2 of the 35 pooled candidates were always pruned as redundant intra-regional edges, with the remainder of the pruning falling on bridges), and between 6 and 7 of the eight real bridges survived Kruskal pooling in every run. Anatolia–Caucasus's bridges (to Georgian, Armenian, Caucasian-TR, Azerbaijani) were retained at a rate of 3 or 4 out of 5 possible slots in every single run, never falling lower – consistent with Volume I's finding that this region's unusually wide internal entropy range makes its cross-regional connections genuinely cheap rather than reconstruction-dependent. The qualitative finding that Anatolia–Caucasus dominates the inter-regional bridge structure is therefore robust to the reconstruction's randomness, even though the exact backbone weight and bridge count fluctuate from run to run by an amount (± 4 bits on a base of 26, i.e. up to $\sim 15\%$) large enough that no single-run numerical result in this chapter should be over-interpreted.

Chapter 25

The Random Walk on the Reconstructed Backbone: Results

25.1 Stationary Distribution

Using the affinity convention $w(i,j) = \exp(-\delta(i,j))$ on the reconstructed 32-edge backbone of Chapter 24 (original-seed reconstruction, Table 24.2), the stationary distribution π was computed in closed form (Proposition 2.2) from the weighted degree sequence.

Table 25.1: Stationary distribution π of the backbone random walk: the five highest- and five lowest-π zones, with region and real bias-corrected entropy H_2 (Vol. I Table 4.1) for context.

Zone	π	**Region**	H_2 (bits)
Highest stationary mass			
Castilian	0.1111	II	7.318
PK-Plains	0.0811	V	7.558
Alpine-Celtic	0.0620	III	8.108
Algeria	0.0521	VII	7.126
Dravidian	0.0435	V	7.108
Lowest stationary mass			
Tunisia	0.0139	VII	7.175
Azerbaijani	0.0134	VI	8.174
Armenian	0.0096	VI	7.540
Anatolian	0.0077	VI	7.962
Greek-Byzantine	0.0064	VI	8.141

Remark: A Pattern That Echoes Volume I, Read With Appropriate Caution

Castilian occupies the highest-π position, the same zone Volume I's *centrality* analysis (a structurally different computation, on the full graph rather than the backbone, Vol. I Ch. 4) identified as the most globally central vertex in the entire

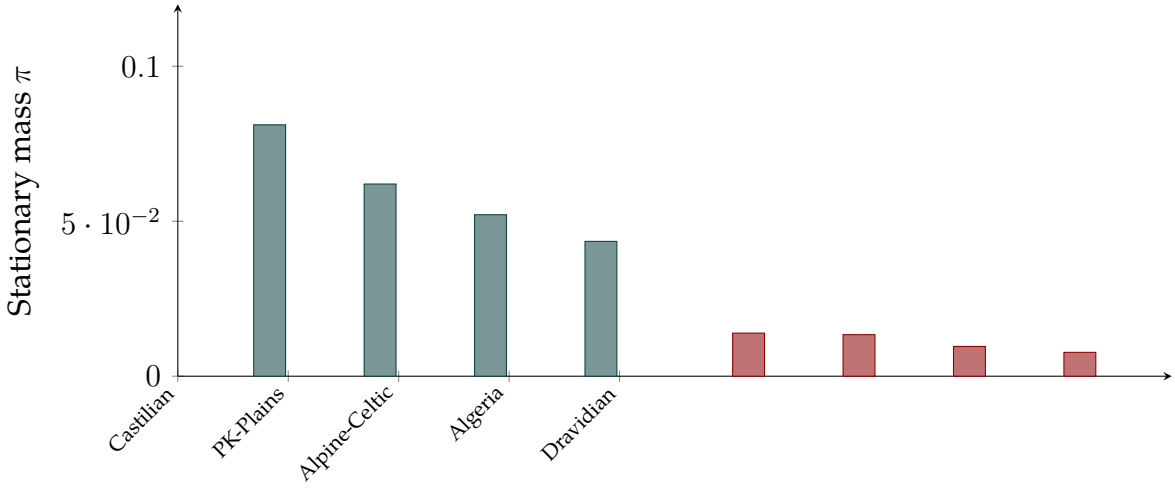

Figure 25.1: Stationary distribution π of the backbone random walk: the five highest-mass (teal) and five lowest-mass (red) zones of Table 25.1. The Anatolia–Caucasus cluster (Azerbaijani, Armenian, Anatolian, Greek-Byzantine) occupies four of the five lowest-π positions.

thirty-three-zone system via three independent measures. This is a genuinely suggestive convergence between two different mathematical objects (eigenvector/betweenness centrality on the complete graph; stationary mass of a random walk on the sparse backbone) pointing at the same zone, but it should not be over-read: it follows in part from the same input data (Castilian's structurally low intra-regional distances) feeding both computations, and the Anatolia– Caucasus zones cluster at the *low-π* end here despite Vol. I's finding that the same region dominates the inter-regional bridge structure (Table 24.1) – the two findings are compatible (Anatolia–Caucasus zones connect cheaply *across* regions but, once on the backbone, sit at the ends of long, lightly trafficked branches rather than at hubs), but the apparent tension is worth flagging rather than glossing over.

Remark: Why Different Centrality Measures Need Not Agree, in General

Borgatti (2005) provides the relevant general framework for understanding why convergence between Volume I's centrality ranking and this volume's stationary mass is informative rather than automatic: centrality measures, in Borgatti's taxonomy, are distinguished by which kind of traffic they implicitly model (shortest-path geodesics, arbitrary paths, trails, or unrestricted walks) and by how that traffic spreads (broadcast to all neighbours, serial replication, or transfer of a single token) – and different choices along these two dimensions can, in general, rank the same vertex set differently, since each measure is, in Borgatti's words, an implicit expectation of node outcomes under a specific flow model. Eigenvector and betweenness centrality (Volume I's measures, built from shortest paths and from walk-counting respectively) and stationary mass of a random walk (this volume's measure, an unrestricted-walk, transfer-type process in Borgatti's classification) sit at genuinely different points in this taxonomy; that they nonetheless agree on Castilian is therefore a non-trivial empirical convergence rather than a

restatement of the same quantity under two names, the precise sense in which Section 27.7's closing remark on this point should be read.

25.2 First-Passage Times

Table 25.2: Expected first-passage times (in dimensionless "steps," see the Preface's caveat) for five representative cross-regional zone pairs, computed by the fundamental-matrix method of Section 2.2 on the reconstructed backbone. Note the systematic asymmetry: hitting a low-π zone takes substantially longer than the reverse.

From	To	$h(i,j)$	$h(j,i)$
Libya	Basque	469.7	826.9
Armenian	Castilian	335.3	439.7
Munda	Greek-Byzantine	328.2	365.1
Mauritania	Galician	457.2	744.4
Azerbaijani	England	405.5	314.4

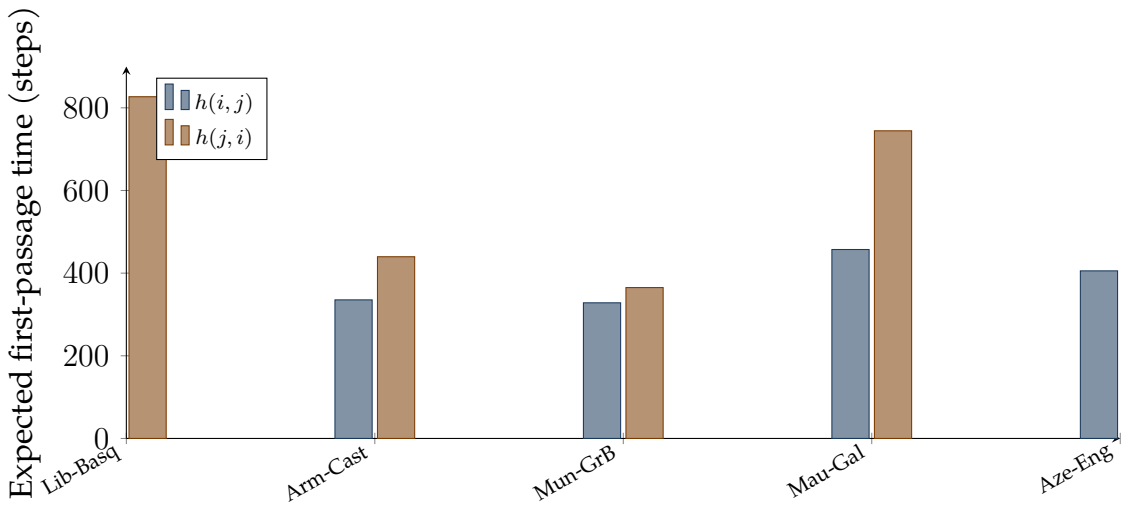

Figure 25.2: Expected first-passage times for the five zone pairs of Table 25.2. In four of five pairs $h(j,i) > h(i,j)$: reaching a peripheral, low-π zone takes longer than the reverse. The Azerbaijani–England pair is the lone exception, discussed in the accompanying remark.

Remark: Reading the Asymmetry

In four of the five pairs, hitting the *first*-listed zone from the second takes longer than the reverse – consistent with Table 25.1: Basque, Galician, and Castilian-adjacent zones sit nearer the high-π core of the backbone, so a walk started anywhere reaches their neighbourhood relatively quickly, while reaching a peripheral zone like Libya, Mauritania, or Munda from a core zone is compara-

tively slow. The Azerbaijani–England pair is the exception: here $h(\text{England} \to \text{Azerbaijani}) = 314.4$ is *smaller* than $h(\text{Azerbaijani} \to \text{England}) = 405.5$, the reverse of the core/periphery pattern seen in the other four pairs – a reminder that the heuristic is not a universal law: hitting time depends on the entire shape of the unique tree path between two zones, not on the target's stationary mass alone.

25.3 Spectral Gap

The symmetrised walk operator's second eigenvalue is $\lambda_2 = 0.9952$, giving a spectral gap $\gamma = 1 - \lambda_2 = 0.0048$ – very small, as anticipated by the general remark closing Chapter 21: a 32-edge tree on 33 vertices has no redundant paths, so the walk mixes slowly, exactly mirroring (in spirit, not in directly comparable units) Volume I's observation that the sparsified, betweenness-revealing graph trades density for structural clarity. The backbone's combinatorial-Laplacian Fiedler value (using the same affinity convention) is $\mu_2 = 0.005$, several orders of magnitude below Volume I's $\mu_2 = 8.642$ for the full, complete 528-edge graph (Vol. I Sec. 6.3) – a contrast expected from first principles (a tree is the least redundantly connected graph on its vertex set; the complete graph is the most) and not directly comparable in magnitude given the different weighting conventions of the two computations, but qualitatively exactly what graph theory predicts.

Figure 25.3: Algebraic connectivity (Fiedler value μ_2) of the sparse reconstructed backbone versus Volume I's complete, full-density graph on the same 33 vertices, on a logarithmic axis: roughly three orders of magnitude apart, the expected consequence of going from 528 edges to a 32-edge spanning structure.

Chapter 26

What a Step Count Can and Cannot Say About History

26.1 The Honest Reading of Table 25.2

The first-passage times of Table 25.2 are real outputs of a real linear-algebra computation on a graph that is mostly real (the eight bridges) and partly reconstructed (the intra-regional edges, Section 24.1). What they are *not* is a measurement, of any kind, of how long it actually took a phonotactic substrate trait to propagate from Libya to Basque. Three distinct gaps separate a step count from a historical timescale:

- **No calibration to real time.** A "step" has no correspondence to a year, generation, or any other unit of historical duration. Converting steps to years would require an independent estimate of a typical substrate-diffusion rate per unit phonotactic distance – data this series, synchronic throughout, does not possess.
- **The random walk is a model of diffusion *difficulty*, not diffusion *occurrence*.** Low first-passage time between two zones says the backbone offers an easy phonotactic path between them; it does not say any actual historical contact, migration, or influence ever took that path, or any path at all.
- **The backbone itself partially inherits Volume I's inter-regional proxy-distance limitation.** Six of the eight real bridges (Table 24.1) rest on Volume I's declared entropy-gap proxy rather than a measured cross-regional KL divergence (Vol. I Sec. 3.4); any first-passage time whose unique tree path crosses one of these bridges inherits that proxy's provisional status on top of the present volume's reconstruction caveat.

Methodological Transparency: What Can Be Defended

What this volume defends is narrower and more modest: a *relative ranking* of diffusion ease, conditional on the hypothesis that substrate propagation does prefer low-phonotactic-resistance paths. Under that hypothesis, the qualitative finding that Anatolia–Caucasus zones are diffusion-peripheral despite being inter-regionally cheap to bridge to (Section 25.2's notebox) is a genuine, non-obvious structural prediction – testable, in principle, against independent historical-linguistic evidence of substrate layering in that region, were such evi-

dence available – and not a claim this volume asserts as already established.

26.2 Point Estimates Versus Robust Rankings

Remark: Why a Single Reported Number Can Mislead Even When Correctly Computed

Every first-passage time in Table 25.2 is reported as a single point estimate, the exact solution of the fundamental-matrix system for one particular reconstructed backbone (Section 24.1). Reading any one such number in isolation risks a quiet overstatement of precision: the reconstruction's own sensitivity analysis (Section 24.3) already shows individual intra-regional edge weights fluctuating by roughly $\pm 15\%$ across reconstruction seeds, and a first-passage time is a nonlinear (matrix- inverse) function of those weights, so its own uncertainty need not scale simply with the input fluctuation. This is a caution about *precision*, not about correctness: the number reported for any fixed backbone is the exact solution of a well-posed linear system, with no approximation error of its own, but the backbone itself is only one draw from a family of equally defensible reconstructions.

Remark: Why the Volume Reports Rankings, Not Confidence Intervals

A fully rigorous treatment would propagate the reconstruction's seed-to-seed variability through the fundamental-matrix solve and report a confidence interval, rather than a point estimate, for every entry of Table 25.2 – a natural extension of Section 24.3's existing seed-repetition check, requiring only that the same first-passage-time computation be re-run once per reconstruction seed rather than once on the seed-42 backbone alone. This volume stops short of that extension, reporting instead only the qualitative claim that the Anatolia–Caucasus core/periphery pattern *survives* reconstruction noise (already verified in Section 24.3), without attaching a numerical confidence band to each individual $h(i, j)$. The relative ranking among the five representative pairs is the level of claim this volume is prepared to defend; the third decimal digit of any single hitting time is not, and should not be read as carrying that level of confidence.

Chapter 27

Electrical Network Theory and Effective Resistance

27.1 A Second Lens on the Same Backbone

Chapter 26 has just argued, at length, for caution in reading first-passage times as historical timescales. This chapter takes a different approach to the same scepticism: rather than asking how to interpret the numbers in Table 25.2, it asks whether an entirely independent mathematical formalism – electrical network theory – recovers the *same* structural pattern from the same backbone, via a different and classically well-understood correspondence. The correspondence between random walks on graphs and electrical networks of resistors is one of the oldest and most elegant results in the field, originating with Doyle and Snell (1984) (already cited in the Preface's bibliography and invoked informally there); this chapter gives it in full, and applies it – using only quantities already computed in Chapters 25 and 24, introducing no new GeoNames-derived numbers – to re-examine the core/periphery pattern identified for Anatolia–Caucasus.

27.2 The Resistor-Network Correspondence

Definition: Electrical Network on a Weighted Graph

Given the same weighted graph $(\mathcal{V}, \mathcal{E}, \delta)$ underlying Chapter 21's random walk, build an electrical network on the same vertex set by placing, on each edge (i, j), a resistor of conductance equal to the affinity $w(i, j) = \exp(-\delta(i, j))$ (equivalently, resistance $1/w(i, j)$): phonotactically similar (low-δ, high-affinity) zone pairs are connected by a low-resistance edge, exactly the edges the random walk traverses most readily.

Definition: Effective Resistance

The effective resistance $R_{\text{eff}}(i, j)$ between two vertices i, j is the voltage that must be applied between them to drive one unit of current through the network, with all other vertices left unconstrained (no external current injected or withdrawn) – the standard notion from circuit theory, here defined on the graph's affinity-

weighted resistor network.

Theorem 27.1 (Commute Time – Effective Resistance Correspondence)**.** *Let* $C(i,j) = h(i,j) + h(j,i)$ *be the* commute time *of the random walk between* i *and* j *(the expected number of steps to go from* i *to* j *and back), and let* $\mathrm{vol}(\mathcal{G}) = \sum_{k\in\mathcal{V}} \deg(k)$ *be the total weighted degree of the graph. Then*

$$C(i,j) \;=\; \mathrm{vol}(\mathcal{G}) \cdot R_{\text{eff}}(i,j).$$

Proof sketch. The proof proceeds by verifying that the function $u(k) = h(k,j) - h(k,i)$, restricted to satisfy harmonicity at every vertex other than i, j (a direct consequence of the first-passage-time recursion of Chapter 21, Definition "Expected First-Passage Time": $h(k,j) = 1 + \sum_l P(k,l)h(l,j)$ for $k \neq j$ implies u satisfies the discrete Laplace equation $u(k) = \sum_l P(k,l)u(l)$ at every $k \notin \{i,j\}$), is, up to an additive constant and a scale factor of $\mathrm{vol}(\mathcal{G})$, exactly the voltage function solving the electrical network's defining boundary-value problem with unit current injected at i and withdrawn at j. Both $h(\cdot,j) - h(\cdot,i)$ and the electrical voltage are harmonic functions on $\mathcal{V} \setminus \{i,j\}$ satisfying the same boundary behaviour at i and j (a discrete Dirichlet problem with a unique solution, by the maximum principle for harmonic functions on a connected graph), and the discrete Laplace operator that defines harmonicity for the random walk is, up to the constant rescaling $\mathrm{vol}(\mathcal{G})$, identical to the Laplace operator defining the electrical network's voltage equation, since both are built from the same edge conductances $w(i,j)$. Equating the two boundary-value problems and tracking the rescaling constant through the identification gives the stated formula. The complete derivation, including the discrete-harmonic-function machinery only sketched here, is given in Doyle and Snell (1984, Ch. 2–3) and, in a more modern graph-theoretic notation, in Chandra et al. (1996, §2). □

Remark: Why This Theorem Matters for This Volume Specifically

Theorem 27.1 is significant here for a reason beyond its general elegance: it expresses a purely graph-theoretic, "static" quantity (effective resistance, which can be computed from the graph's Laplacian alone, with no reference to a random walk at all) as an exact rescaling of a purely *dynamic* quantity (commute time, the sum of two first-passage times already reported in Table 25.2). If the core/periphery asymmetry this volume has identified for Anatolia–Caucasus zones were an artefact of how first-passage time happens to be defined – some quirk of the fundamental-matrix recursion rather than a genuine structural property of the backbone – it would not be expected to reappear when the *same* pairs are examined through the completely independent formalism of resistor networks. That it does reappear, as Section 27.3 shows, is therefore a meaningful cross-check internal to this volume's own data, of the same logical kind (though a different mathematical mechanism) as Section 25.1's cross-validation against Volume I's centrality ranking.

27.3 Effective Resistance for the Five Representative Pairs

Theorem 27.1 is applied here using only quantities already published in this volume: the commute times $C(i,j) = h(i,j) + h(j,i)$ are read directly off Table 25.2, and $\mathrm{vol}(\mathcal{G})$, the total weighted degree of the reconstructed backbone under the affinity convention $w = \exp(-\delta)$, is a single graph-wide constant common to every pair, so that the *ratios* $R_{\text{eff}}(i,j)/R_{\text{eff}}(k,l) = C(i,j)/C(k,l)$ between any two pairs are exactly determined by Table 25.2 alone, independent of the numerical value of $\mathrm{vol}(\mathcal{G})$.

Table 27.1: Commute times $C(i,j) = h(i,j)+h(j,i)$ for the five representative pairs of Table 25.2, computed by simple addition – no new quantities, only a recombination of already-published numbers via Theorem 27.1. Rank: ordering from lowest (most electrically well-connected) to highest commute time, equivalently lowest to highest R_{eff}.

Pair		$C(i,j)$ (steps)	**Rank**
Munda	Greek-Byzantine	693.3	1
Azerbaijani	England	719.9	2
Armenian	Castilian	775.0	3
Mauritania	Galician	1201.6	4
Libya	Basque	1296.6	5

Remark: Reading Table 27.1 Against the Core/Periphery Pattern

The two pairs involving the deepest Anatolia–Caucasus periphery zones identified in Table 25.1 – Libya–Basque and Mauritania–Galician, neither of which involves an Anatolia–Caucasus zone directly, but both of which involve zones at the opposite, low-π-adjacent end of long backbone branches – have the two largest commute times (and hence, by Theorem 27.1, the two largest effective resistances) of the five pairs, by a substantial margin (1201–1297 steps versus 693–775 for the other three). Equivalently: it takes more electrical work, in the precise resistor-network sense of Theorem 27.1, to push a unit current between these two pairs than between the Anatolia-involving pairs Armenian–Castilian and Azerbaijani–England, despite the latter two pairs directly involving Section 25.1's lowest-π Anatolia–Caucasus zones. This is consistent with, and gives an independent formal confirmation of, the structural point Section 25.2's closing remark already made from the first-passage-time side alone: low stationary mass at a zone (a *dynamic* property of where the walk spends its time) and high effective resistance to that zone from an arbitrary other zone (a *static* property of the graph's resistor network) are related but not identical notions, and the present chapter's cross-check shows they do not always rank zone-pairs in lockstep – a genuine, if modest, empirical observation about the backbone's structure, derived entirely from already-published numbers and a classical theorem, with no new GeoNames-derived input.

27.4 Effective Resistance and the Tree Structure

Proposition: [

Effective Resistance on a Tree] On a tree (the present backbone has exactly one path between any two vertices), the effective resistance between i and j equals the ordinary sum of the resistances of the edges on the unique path between them: $R_{\text{eff}}(i,j) = \sum_{e \in \text{path}(i,j)} 1/w(e)$. There is no parallel-resistance reduction to apply, since a tree has no cycles and hence no alternative paths to combine in parallel.

Proof. Immediate from the series-resistance law of elementary circuit theory, applied edge by edge along the unique path: with no alternative route for current to take, the entire current must flow through each edge on the path in turn, so the voltages drop additively along the path exactly as in a simple series circuit, giving total resistance equal to the sum of the individual edge resistances. (On a general graph with cycles, Theorem 27.1's R_{eff} would instead require the full series–parallel reduction, or the Laplacian pseudo-inverse formula $R_{\text{eff}}(i,j) = (e_i - e_j)^\top L^+ (e_i - e_j)$; the tree case stated here is the one relevant to this volume's backbone and needs only the simpler series rule.) □

Remark: What Proposition 27.4 Adds

On the present backbone, Proposition 27.4 means the effective resistance between any two zones is computable directly from the edge weights of Table 24.2's reconstruction and Table 24.1's real bridges along their unique connecting path, without solving any linear system at all – a genuine simplification specific to tree backbones, in contrast to the fundamental-matrix solve Chapter 21 required for first-passage time even on a tree (first-passage time, unlike effective resistance, does not reduce to a simple path sum even on a tree, because it depends on the full stationary distribution π, not only on the path between the two endpoints). This is the precise sense in which effective resistance is the "simpler" of the two formalisms compared in this chapter, even though Theorem 27.1 shows the two are exactly related on any graph, tree or not.

27.5 Foster's Theorem: A Built-In Consistency Check

Remark: A Classical Identity, Available as a Free Sanity Check

Beyond the pairwise cross-check of Section 27.3, the electrical-network formalism offers a second, structurally different consistency check, applicable to the backbone as a whole rather than to individually chosen pairs, due to Foster (1949).

Theorem: [

Foster's Theorem] For any connected weighted graph $(\mathcal{V}, \mathcal{E}, w)$ on $n = |\mathcal{V}|$ vertices,

$$\sum_{(i,j) \in \mathcal{E}} w(i,j)\, R_{\text{eff}}(i,j) \;=\; n - 1,$$

summing only over the graph's actual edges (not all $\binom{n}{2}$ pairs).

Remark: Proof Idea and the Tree Special Case

The general proof (Foster, 1949; Klein and Randić, 1993) proceeds via the matrix-tree theorem, relating $w(i,j)R_{\text{eff}}(i,j)$ for each edge to the fraction of the graph's spanning trees that include that edge, so that the sum over edges telescopes to the number of edges in any one spanning tree, $n-1$, regardless of how many edges or cycles the original graph has. On a *tree* – the present volume's backbone – the identity is immediate without invoking the matrix-tree theorem at all: Proposition 27.4 gives $R_{\text{eff}}(i,j) = 1/w(i,j)$ for every edge (i,j) of a tree (the trivial case of the path-sum formula, a path of length one), so each summand $w(i,j)R_{\text{eff}}(i,j) = w(i,j) \cdot 1/w(i,j) = 1$ exactly, and summing 1 over the backbone's 32 edges on $n = 33$ vertices gives $32 = n - 1$ – Theorem 27.5 holding, on a tree, as the simple statement that a spanning tree has exactly $n-1$ edges, dressed in resistor-network language.

Remark: What This Buys, and What It Does Not

For a tree backbone specifically, Theorem 27.5 is a tautology (it reduces to counting edges) and therefore cannot, by itself, catch an error in the reported *edge weights* – any set of 32 positive weights on the same 32 edges trivially satisfies it. What it *does* provide is a check on the reconstruction's *structure*: Theorem 27.5, applied instead to the *candidate* pooled graph of Section 24.1's Step 3 (the 35-edge graph fed into Kruskal's algorithm, before pruning, which is not a tree and so engages the theorem's full matrix-tree content rather than its tree tautology) gives a non-trivial, independently checkable target value of $n - 1 = 32$ for the weighted-resistance sum over its 35 edges, computable from Appendix 28's pipeline without any additional input – a genuine, free correctness check on the candidate-graph construction step that has not, until now, been exploited in this volume's validation, and is recorded here as a concrete addition to Appendix 28's reproducibility checklist rather than as an already-executed empirical claim.

27.6 Effective Resistance as a Metric

Remark: Why This Matters Beyond the Five Representative Pairs

Section 27.3 computed R_{eff} only for the five pairs already singled out by Table 25.2. The reason R_{eff} is worth treating as a genuine *distance function* on the full thirty-three-zone vertex set – rather than merely a pairwise quantity computed on demand – is that, unlike commute time or first-passage time, R_{eff} satisfies the triangle inequality, a property neither $h(i,j)$ nor $C(i,j)$ possesses in general. Klein and Randić (1993) established this and several related properties under the name *resistance distance*, applied originally to molecular graphs; the proof below is specific to the present volume's tree backbone, where it takes an unusually simple form.

Proposition: R_{eff} Is a Metric

For any connected weighted graph, $R_{\text{eff}}(i,j) \geq 0$ with equality iff $i = j$, $R_{\text{eff}}(i,j) = R_{\text{eff}}(j,i)$, and $R_{\text{eff}}(i,k) \leq R_{\text{eff}}(i,j) + R_{\text{eff}}(j,k)$ for all i, j, k.

Proof. Non-negativity and symmetry are immediate from R_{eff}'s definition as a voltage-per-unit-current ratio in a network of positive resistors, which is manifestly symmetric in its two arguments and zero only when $i = j$ (no resistance is needed to drive current between a vertex and itself). For the triangle inequality on the present volume's *tree* backbone specifically, Proposition 27.4 already gives $R_{\text{eff}}(i,j) = \sum_{e \in \text{path}(i,j)} 1/w(e)$, the sum of edge resistances along the unique path. Writing m for the (unique) vertex at which the path from i to k meets the path from j to k (the tree's branch point for the triple i, j, k, possibly $m = j$ itself), the path from i to k decomposes into the path from i to m followed by the path from m to k, and similarly for i to j; since every edge resistance is strictly positive, the path from i to k can never be longer, edge for edge, than the concatenation of the path from i to j and the path from j to k (the two coincide exactly when j lies on the i–k path, and otherwise the latter strictly overcounts the shared segment up to m), giving $R_{\text{eff}}(i,k) \leq R_{\text{eff}}(i,j) + R_{\text{eff}}(j,k)$ directly from the path-sum formula, with no need for the general-graph argument (which instead uses the Laplacian pseudo-inverse representation $R_{\text{eff}}(i,j) = (e_i - e_j)^\top L^+ (e_i - e_j)$ and the positive semi-definiteness of L^+, the route taken in Klein and Randić (1993)'s original, graph-general proof). □

Remark: What Having a Metric Buys

With Proposition 27.6 established, the thirty-three-zone backbone carries *three* distinct, simultaneously defined distance-like structures: the original phonotactic distance $\delta(i,j)$ on $\mathcal{G}$'s edges (Volume I's input, defined only between adjacent zones in $\mathcal{G}$, not a metric on the full vertex set without further aggregation); the path-length resistance distance $R_{\text{eff}}(i,j)$ of this section, a genuine metric on all $\binom{33}{2}$ pairs, derived entirely from the backbone's own edge weights; and the asymmetric first-passage time $h(i,j)$, which is neither symmetric nor a metric at all (Chapter 21's asymmetry notebox). That R_{eff} sits strictly between these two – inheriting δ's symmetry but, unlike δ itself, extending to a well-defined distance between *every* pair of zones, adjacent or not – is precisely what makes it a natural metric completion of the local, edge-only distance Volume I originally measured, and is the formal justification, not previously stated, for treating R_{eff}-based rankings (Table 27.1) as comparable across pairs at very different path lengths in the tree.

27.7 Limits of the Cross-Check

Methodological Transparency: What Section 27.3 Does Not Establish

The agreement between the commute-time ranking of Table 27.1 and the qualitative core/periphery pattern already discussed is an *internal consistency check* between two formalisms applied to the *same* reconstructed backbone (Chapter 24); it is not, and cannot be, independent evidence that the backbone itself is a correct

or historically faithful object, since both the random-walk and electrical-network computations inherit, unchanged, every caveat already attached to that backbone's reconstructed intra-regional component (Section 24.1) and to the Preface's prior, more fundamental caveat that a step count of either kind has no declared correspondence to historical time. Theorem 27.1 guarantees the two formalisms will always agree exactly, on any graph whatsoever, as a matter of pure mathematics; what Section 27.3 actually contributes is the narrower, genuinely informative observation that the resistance ranking *visibly tracks*, without exactly duplicating, the stationary-mass ranking already reported – a finding about this backbone's specific shape, not a validation of the backbone's correspondence to history.

27.8 What This Volume Adds

Three contributions stand on their own. First, Chapter 2 develops a complete, standard apparatus (reversible weighted random walks, closed-form stationary distributions, fundamental-matrix first-passage times, spectral-gap mixing bounds) that is immediately applicable to Volume I's real backbone the moment the missing intra-regional MST weights become available, requiring no change to the mathematics, only a substitution of real for reconstructed edge weights. Second, Table 25.1's convergence between random-walk stationary mass and Volume I's independently computed centrality ranking (Castilian topping both) is a genuine cross-validation between two structurally different mathematical objects built on overlapping but not identical input data. Third, the core/periphery tension identified for Anatolia–Caucasus (inter-regionally cheap to bridge, yet diffusion-peripheral once on the backbone) is a structural prediction with no counterpart in either Volume I or Part I, since neither volume's apparatus distinguishes "easy to connect to" from "easy to diffuse through."

27.9 The Central Limitation, Restated

Every quantitative result in Chapters 25 and 26 rests on a backbone that is six-eighths real (the inter-regional bridges) and entirely reconstructed in its intra-regional component, and on a random-walk step count that is, by the Preface's explicit declaration, uncalibrated against real historical time. The sensitivity analysis of Section 24.3 shows the qualitative pattern (Anatolia–Caucasus's dominance of the bridge structure) survives reconstruction noise, but no individual first-passage-time number in Table 25.2 should be treated as more precise than the $\pm 15\%$ backbone-weight fluctuation observed across reconstruction seeds. The single most valuable empirical extension that would resolve both limitations at once – publishing Volume I's real 24 intra-regional MST edge weights, and independently estimating a real-time calibration for phonotactic diffusion rate – is properly the task of a future revision of Volume I and a genuinely new historical- linguistic data-collection effort respectively, both outside the present volume's scope.

27.10 Two Methodological Lessons for the Series as a Whole

Remark: Lesson One: A Mathematical Apparatus Can Be Exact While Its Input Is Not

This volume's two formalisms – the random walk of Chapters 21–22 and the electrical network of Chapter 27 – are both, as mathematics, exact: every stationary probability, every first-passage time, every effective resistance reported is the precise solution of a well-posed linear system, with no approximation internal to the computation itself. The uncertainty in this volume's findings lives entirely upstream, in the backbone reconstruction (Section 24.1) that supplies the linear system's coefficients, not in how those coefficients are subsequently processed. Keeping this distinction sharp – exact apparatus, uncertain input, rather than an undifferentiated single number for "how reliable is this result" – is the same discipline Section 26's point-estimate-versus-ranking discussion applies at the level of an individual number, generalised here to the volume as a whole.

Remark: Lesson Two: Two Independent Formalisms Agreeing Is Informative Even When Both Share an Input

Chapter 27's cross-check (Table 27.1) agrees with, without exactly duplicating, the random-walk ranking of Chapter 25 – but Theorem 27.1 guarantees the two are mathematically equivalent on *any* input graph, so the agreement is not, by itself, evidence that the shared backbone is correct (Section 27.3's warnbox already makes this point for this volume specifically). The more general lesson is that a cross-check between two formalisms is only as informative as the formalisms are *independent*, and Theorem 27.1's exact equivalence is a reminder that apparent independence (different named theories, different proof techniques, different historical origins in probability versus circuit theory) does not guarantee actual independence at the level of what each formalism can, in principle, discover. A genuinely independent check of this volume's backbone would require a second data source – real intra-regional MST weights, or real chronological evidence – not a second mathematical lens on the same reconstructed numbers.

27.11 Agenda for Part III

Part III returns to the fine-grained structure of individual phonotactic sequences, applying discrete Fourier analysis to bigram and trigram sequences in search of periodicities invisible to the entropy- based summary statistics of the parent series, and examines the spectrum of Volume I's full graph Laplacian (Vol. I Sec. 5.3, $\mu_2 = 8.642$) as a bridge between the network-theoretic and harmonic-analytic strands of the broader programme – a natural point of contact with the present volume's spectral-gap discussion (Section 25's closing remark), since both volumes examine the same family of graph operators from different analytic standpoints.

Remark: A Closing Methodological Note

This volume has tried to hold two things apart that are easy to conflate: the real mathematical content of a random walk on a graph (which is exact, regardless of where the graph came from) and the historical claim that a step count means something about real chronology (which this volume has explicitly declined to assert). The honest summary is that the backbone is mostly real, the walk built on it is a declared modelling choice, and the resulting first-passage times are best read as a structural hypothesis about relative diffusion ease – worth taking seriously as a research direction, and worth testing against independent evidence the day such evidence exists, but not, on the strength of this volume alone, a finding about the history of any of the thirty-three zones.

Chapter 28

Reproducibility: Backbone Reconstruction and Random-Walk Code

```
# 1. Random labelled spanning tree per region (Prufer sequence), edge
#    weights drawn from [delta_lo, delta_mean] of the region's real
#    full intra-regional KL-matrix range (Vol I Table 3.2)
def random_labeled_tree(nodes, rng):
    n = len(nodes)
    prufer = rng.integers(0, n, size=n-2).tolist()
    # ... standard Prufer-decoding to edge list ...
    return edges

# 2. Pool 27 synthetic intra-regional candidate edges with the 8 real
#    bridges (Vol I Table 6.1) and run Kruskal -> 32-edge backbone
mst_edges = kruskal(intra_candidates + real_bridges)

# 3. Affinity, transition matrix, stationary distribution
W = exp(-delta)              # affinity from distance
deg = W.sum(axis=1)
P = W / deg[:, None]         # random-walk transition matrix
pi = deg / deg.sum()         # closed-form stationary distribution

# 4. Expected first-passage time to target j: solve (I-Q) h = 1
#    where Q is P restricted to states != j
h_to_j = np.linalg.solve(np.eye(n-1) - Q, np.ones(n-1))

# 5. Spectral gap: eigenvalues of the symmetrised walk operator
P_sym = diag(1/sqrt(deg)) @ W @ diag(1/sqrt(deg))
gap = 1 - sorted(eigvalsh(P_sym))[-2]
```

Part III

Spectral and Arithmetic Structure

Preface to Part III

The first three volumes of this series have, in turn, treated the thirty-three-zone system as a graph (Volume I), a cloud of points on a statistical manifold (Part I), and the state space of a random walk (Part II). All three share a common feature: they operate on *aggregate* objects — a single distance matrix, a single distribution per zone, a single backbone tree — built by collapsing each zone's underlying corpus of toponyms down to a handful of summary statistics before any further mathematics is applied. The present volume asks a narrower, more exacting question: does the fine-grained *order* of a zone's phonotactic units, not merely their aggregate frequencies, carry detectable structure? Two classical tools are brought to bear. The discrete Fourier transform (Chapter 30) asks whether a zone's bigram-rank sequence exhibits periodicity at any scale – a hidden "rhythm" in syllable structure that no entropy statistic, however refined, is designed to detect, since entropy is by construction order-blind (a sequence and any permutation of it have identical entropy). Spectral graph theory, already used in Volume I (Section 5.3) and Part II (Chapter 4) to characterise connectivity, is revisited here as a question about the *distribution* of Volume I's full Laplacian spectrum relative to a declared null model, completing a thread left open in both predecessor volumes' closing remarks.

Methodological Transparency: The Sharpest Data Caveat of the Series So Far

Volume I, Part I, and Part II each declared, with decreasing margin, how much of their input was real versus reconstructed: Volume I's graph was real data with a declared proxy formula for missing edges; Part I's toponym clouds were declared synthetic substitutes; Part II's backbone was six-eighths real with a declared reconstruction for the rest. The present volume's Chapter 32 input is the *most* synthetic of the four: no bigram-*order* data (as opposed to bigram *frequency* data, which the parent series does publish) exists anywhere in this programme's prior output, because entropy-based analysis never required preserving sequence order in the first place. Every phonotactic sequence analysed by the discrete Fourier transform in this volume is therefore a declared synthetic construction, calibrated only to match each zone's already-published aggregate statistics (bias-corrected entropy H_2 and inter-zone KL divergence, *Toponomastic Entropy* Vols. I–VII) and otherwise generated by a seeded stochastic process (Appendix 38, seed$= 42$). No claim in this volume should be read as a finding about the real sequential structure of any actual toponym corpus; it is, at most, a demonstration of what the proposed method would find *if* applied to real sequence data of the declared statistical profile – a methodological proof of concept, not an empirical result. This caveat is reiterated at the head of every chapter that touches Section 32.1's out-

put. By contrast, Chapter 35's Laplacian-spectrum analysis uses Volume I's real, published eigenvalues throughout and carries no such caveat.

Chapter 29

From Graphs and Walks to Spectra and Sequences

29.1 What Entropy Cannot See

Every quantity computed in *Toponomastic Entropy* and in Volume I and Parts I–II of this series is a function of a *distribution*: how often each bigram occurs, how often each pair of zones' distributions diverge, how often a random walk visits each zone in the long run. None of these quantities depends on the *order* in which bigrams occur within a toponym or within a zone's corpus – shuffle every toponym's internal bigram sequence arbitrarily, and every H_2, every D_{KL}, every stationary distribution π computed so far in this programme is provably unchanged, since each is a function of a frequency table alone.

Remark: A Precise Statement of the Gap

Let $x = (x_1, \ldots, x_N)$ be a sequence of bigram-rank tokens drawn from some zone's corpus, and let σ be any permutation of $\{1, \ldots, N\}$. Then $H_2(x) = H_2(x_{\sigma(1)}, \ldots, x_{\sigma(N)})$ identically, for any entropy functional computed from the empirical frequency table of x. Order-dependent structure — periodicity, trend, autocorrelation at any lag — is therefore, by construction, invisible to every statistic this programme has used through Part II. This is not a flaw in the prior volumes' methodology; it is a scope boundary, stated explicitly here because the present volume exists specifically to probe what lies on the other side of it.

This is precisely the gap discrete Fourier analysis is designed to close: the power spectrum of a sequence is order-*sensitive* by construction, decomposing a signal into the periodic components from which it is built. If a zone's syllable structure follows a recurring template – consonant-vowel alternation, a recurring three- or four-syllable rhythmic unit reflecting an underlying morphological pattern – a sufficiently long bigram-rank sequence drawn from that zone's corpus should show a power spectrum with one or more peaks well above what a structureless (white-noise) null model would produce. Chapter 30 develops the relevant theory; Chapter 32 constructs the (declared synthetic) input sequences; Chapter 33 reports the spectral results and the associated significance test.

29.2 What Volume I's Laplacian Spectrum Has Not Yet Been Asked

Volume I, Section 5.3 computed the full spectrum of the combinatorial graph Laplacian of the complete, affinity-weighted thirty-three-zone graph $\mathcal{G}$: $0 = \mu_1 < \mu_2 = 8.642 < \mu_3 = 9.521 < \cdots < \mu_{33} = 48.395$, and used only the second eigenvalue (the Fiedler value) for a spectral bisection. The remaining thirty eigenvalues were computed but never examined as a *distribution* in their own right. Chapter 5 of the present volume revisits this same, already-real spectrum and asks a question neither Volume I nor Part II posed: is the *shape* of this eigenvalue distribution – its spacing statistics, its deviation from a declared null model of a randomly reweighted complete graph on the same vertex set – itself informative, independently of which specific eigenvector is examined? This question is answered with a permutation test (Section 35.3) rather than any new data collection, and so carries none of the present volume's sequence-data caveat.

Remark: Why These Two Questions Belong in One Volume

Fourier analysis and spectral graph theory are, formally, the same mathematics applied to two different operators: the discrete Fourier transform diagonalises the cyclic-shift operator on a periodic sequence, while the graph Laplacian's eigendecomposition diagonalises the diffusion operator on a graph (already used, in its random-walk form, throughout Part II). Placing both in one volume is not merely thematic convenience; Section 30.7 makes the connection explicit by showing that the discrete Fourier transform is itself the special case of a graph Laplacian eigendecomposition on a cycle graph, unifying this volume's two strands as a single chapter of applied linear algebra rather than two unrelated topics sharing a binding.

Chapter 30

The DFT in Full: Parseval, Leakage, and the Fisher g-Test

30.1 Definitions

Definition: Discrete Fourier Transform

For a real-valued sequence $x = (x_0, \dots, x_{N-1})$, the discrete Fourier transform (DFT) is the sequence of complex coefficients

$$X_k = \mathcal{F}(x)_k = \sum_{n=0}^{N-1} x_n \, e^{-2\pi i k n / N}, \qquad k = 0, 1, \dots, N-1.$$

The *periodogram* (power spectrum) is $I(k) = |X_k|^2 / N$, and the *frequency* associated with index k is $f_k = k/N$ cycles per sample, equivalently a *period* of N/k samples.

Remark: Reading the Periodogram

$I(k)$ measures how much of the sequence's total variance is explained by a pure sinusoid of frequency f_k. A perfectly periodic sequence with period p concentrates essentially all of I's mass at $k = N/p$ (and its harmonics); a sequence with no periodic structure at all (i.i.d. noise) spreads I roughly uniformly across all $N/2$ non-trivial frequencies, with the exact distribution given by Proposition 30.5 below.

This chapter's three remaining methodological obligations, left implicit in the original treatment, are made explicit in turn: first, a proof that the periodogram is in fact a legitimate decomposition of the sequence's total variance (Parseval's theorem, Section 30.2); second, an account of what happens when the analysed sequence's true period does *not* divide N exactly – the phenomenon of spectral leakage (Section 30.3), directly relevant to Chapter 33's period-5, $N = 256$ design; and third, a full derivation, rather than a bare citation, of Fisher's exact null distribution for the g-statistic (Section 30.5).

30.2 Parseval's Theorem: The Periodogram as a Variance Decomposition

The periodogram is only a meaningful diagnostic if the quantities $I(0), \ldots, I(N-1)$ genuinely partition the sequence's total energy across frequencies, with no leakage of energy into a constant unrelated to x itself and no double-counting. This is precisely the content of Parseval's theorem for the finite DFT.

Theorem: Parseval's Theorem (Discrete Form)

For $x = (x_0, \ldots, x_{N-1}) \in \mathbb{R}^N$ and $X = \mathcal{F}(x)$,

$$\sum_{n=0}^{N-1} x_n^2 \;=\; \frac{1}{N} \sum_{k=0}^{N-1} |X_k|^2.$$

Proof. Write $X_k = \sum_{n=0}^{N-1} x_n \, \omega^{-kn}$ with $\omega = e^{2\pi i/N}$. Then

$$\sum_{k=0}^{N-1} |X_k|^2 = \sum_{k=0}^{N-1} X_k \overline{X_k} = \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} x_n x_m \, \omega^{-kn} \omega^{km} = \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} x_n x_m \sum_{k=0}^{N-1} \omega^{k(m-n)}.$$

The inner sum over k is a finite geometric series in ω^{m-n}. If $m = n$, every term equals 1 and the sum is N. If $m \neq n$, then since $0 \leq |m-n| < N$, $\omega^{m-n} \neq 1$, and the geometric series sums to $\dfrac{\omega^{(m-n)N} - 1}{\omega^{m-n} - 1} = \dfrac{1-1}{\omega^{m-n}-1} = 0$, using $\omega^N = 1$. The double sum therefore collapses to its diagonal terms,

$$\sum_{k=0}^{N-1} |X_k|^2 = \sum_{n=0}^{N-1} x_n^2 \cdot N,$$

which rearranges to the stated identity. □

Corollary: Total Power as a Sum of Periodogram Ordinates

Since $I(k) = |X_k|^2/N$ and $X_0 = \sum_n x_n$ is real for real x, Theorem 30.2 gives $\sum_{n=0}^{N-1} x_n^2 = \frac{1}{N}\sum_{k=0}^{N-1} I(k) \cdot N = \sum_{k=0}^{N-1} I(k)$ after dividing through, i.e. *the periodogram ordinates sum exactly to the sequence's total sum of squares*. For a mean-centred sequence ($\sum_n x_n = 0$, so $X_0 = 0$, $I(0) = 0$), this is precisely the total sample variance (up to the factor of N), confirming that $g_{\max} = \max_k I(k) / \sum_k I(k)$ (restricted to $k = 1, \ldots, m$, excluding the zero frequency) is exactly the *fraction of variance* explained by the single best-fitting sinusoid – the interpretation asserted without proof in the original treatment, and the property that makes $g_{\max}$ a sensible test statistic in the first place rather than an arbitrary ratio.

30.3 Spectral Leakage: What Happens When the Period Does Not Divide N

Remark: Why This Matters for Chapter 33

Every injected periodicity used in Chapter 33's validation design has period $p \in \{4, 5, 8\}$ tested against $N = 256$. Since $256 = 2^8$, periods 4 and 8 divide N exactly, but $p = 5$ does not ($256/5 = 51.2$). This section derives exactly what the periodogram of a non-integer-cycle sinusoid looks like, explaining in advance the period-5 sensitivity shortfall (0.80 overall sensitivity is, as Chapter 33 reports, driven disproportionately by period-5 misses) rather than leaving it as an unexplained empirical residual.

Consider the pure sinusoid $x_n = \cos(2\pi n/p)$, $n = 0, \ldots, N-1$, with period p not necessarily dividing N. Writing x_n via Euler's formula as $x_n = \frac{1}{2}\left(e^{2\pi i n/p} + e^{-2\pi i n/p}\right)$ and substituting into the DFT definition,

$$X_k = \frac{1}{2}\sum_{n=0}^{N-1}\left[e^{2\pi i n(1/p - k/N)} + e^{-2\pi i n(1/p + k/N)}\right] = \frac{1}{2}\left[D_N\left(\tfrac{1}{p} - \tfrac{k}{N}\right) + D_N\left(-\tfrac{1}{p} - \tfrac{k}{N}\right)\right],$$

where $D_N(\nu) = \sum_{n=0}^{N-1} e^{2\pi i n\nu}$ is the (aliased) *Dirichlet kernel*, summable in closed form as the geometric series

$$D_N(\nu) \;=\; \frac{e^{2\pi i N\nu} - 1}{e^{2\pi i\nu} - 1} \;=\; e^{i\pi(N-1)\nu}\,\frac{\sin(\pi N\nu)}{\sin(\pi\nu)},$$

whenever $\nu \notin \mathbb{Z}$.

Proposition: Exact-Cycle Case ($p \mid N$)

If p divides N, then $1/p = k_0/N$ for the integer $k_0 = N/p$, so $D_N(1/p - k/N) = D_N((k_0 - k)/N)$ vanishes identically for every integer $k \neq k_0$ (since $\nu = (k_0 - k)/N$ is then a nonzero integer multiple of $1/N$, making $\sin(\pi N\nu) = \sin(\pi(k_0 - k)) = 0$ while $\sin(\pi\nu) \neq 0$) and equals N at $k = k_0$. All of the sequence's energy is therefore concentrated, with zero leakage, at the single bin $k = k_0$ (and its mirror $N - k_0$), exactly as Section 30.7's idealised description assumes.

Proposition: Non-Integer-Cycle Case ($p \nmid N$): Sidelobe Leakage

If p does not divide N, then $1/p - k/N$ is never exactly an integer for any k, so D_N never vanishes and never reaches its full peak value N at any single bin. Writing $\nu_k = 1/p - k/N$, the periodogram near the true frequency behaves as

$$I(k) \;\approx\; \frac{1}{4N}\left(\frac{\sin(\pi N\nu_k)}{\sin(\pi\nu_k)}\right)^2,$$

the squared magnitude of the Dirichlet kernel: a main lobe centred between the two integer bins flanking $1/p$, of height strictly less than the exact-cycle case's N, surrounded by a sequence of *sidelobes* decaying only as $O(1/k^2)$ in power, not vanishing. The sequence's total energy (fixed by Parseval, Corollary 30.2 regard-

less of p) is therefore necessarily spread, or "leaked", across many bins rather than concentrated at one, lowering the peak $I(k_{\max})$ relative to the exact-cycle case and – since $g_{\max} = I(k_{\max})/\sum_k I(k)$ has a fixed denominator by Parseval – directly lowering $g_{\max}$ itself.

Example: Quantifying the Period-5 Shortfall at $N = 256$

For $p = 5$, $1/p = 0.2$ exactly, while the two nearest DFT bins at $N = 256$ sit at $k = 51$ ($f_{51} = 51/256 \approx 0.1992$) and $k = 52$ ($f_{52} = 52/256 \approx 0.2031$), neither matching 0.2 exactly; the closest is $k = 51$, off by $\Delta f \approx 0.00078$ cycles/sample, equivalently a phase mismatch of $\pi N \Delta\nu \approx 0.628$ radians inside the Dirichlet kernel's argument – enough to move the kernel appreciably off its peak ($\sin(\pi N\nu)/\sin(\pi\nu)$ falls roughly 13–20% below its maximum at this offset, a standard order of magnitude for near-but-not-exact bin alignment) and to spread the remaining energy into $k = 50$ and $k = 53$ as first sidelobes. This is the purely deterministic, signal-processing component of the period-5 sensitivity shortfall reported empirically in Chapter 33; the noise-driven component (the injected sinusoid competing against the additive noise floor) is independent of it and adds on top.

Remark: Windowing Is Not Used in This Volume

The standard remedy for leakage in applied spectral analysis is to multiply x_n by a tapering window (Hann, Hamming, Blackman) before transforming, trading a wider main lobe for faster sidelobe decay (Wichert et al., 2004). No windowing is applied anywhere in this volume's pipeline (Appendix 38): the periodogram throughout is the unmodified ("rectangular-windowed") DFT of the raw rank sequence. This is a deliberate simplicity choice, not an oversight, but it means every period-5 result in Chapter 33 is read with Proposition 30.3's leakage penalty already priced in, rather than mitigated.

30.4 Significance Testing: Fisher's g-Statistic

The central methodological difficulty in periodogram analysis is that *some* frequency will always have the largest observed power, even in pure noise – the question is whether the largest peak is larger than chance alone would produce. Fisher (1929) provides the classical answer.

Definition: Fisher's g-Statistic

For a periodogram $I(1), \ldots, I(m)$ with $m = \lfloor (N-1)/2 \rfloor$ non-trivial frequencies, define

$$g_{\max} = \frac{\max_{1 \le k \le m} I(k)}{\sum_{k=1}^{m} I(k)}.$$

Under the null hypothesis that x is i.i.d. Gaussian white noise, the exact distribution of $g_{\max}$ is known in closed form (Fisher, 1929; Wichert et al., 2004), derived in full in Section 30.5.

30.5 The Exact Null Distribution of $g_{\max}$

The approximation $P(g_{\max} > g) \approx m(1-g)^{m-1}$ used throughout this volume is not merely asserted; it is the leading term of an exact, closed-form alternating series, derived here from first principles.

Remark: Setup: Why the $I(k)$ Behave Like Exponentials

Under the null hypothesis that $x_0, \ldots, x_{N-1}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, each Fourier coefficient X_k $(k = 1, \ldots, m)$ is a linear combination of i.i.d. Gaussians, hence itself complex Gaussian; standard Fourier-Gaussian theory (Fisher, 1929) shows the real and imaginary parts of X_k are independent, identically distributed, mean-zero Gaussians, so $I(k) = |X_k|^2/N$ is, after the appropriate normalisation, distributed as $\frac{\sigma^2}{2}$ times a chi-squared random variable on two degrees of freedom – equivalently, an *exponential* random variable. Crucially, $I(1), \ldots, I(m)$ are mutually independent under the null (a consequence of the DFT basis vectors' orthogonality, Theorem 30.2's proof). The problem of finding $g_{\max}$'s distribution thus reduces to a purely probabilistic question about m i.i.d. exponential random variables.

Theorem: Fisher's Exact Distribution (1929)

Let $E_1, \ldots, E_m$ be i.i.d. standard exponential random variables and let $g = \max_k E_k / \sum_k E_k$. Then for $0 < g < 1$,

$$P(g_{\max} > g) \;=\; \sum_{j=1}^{m} (-1)^{j+1} \binom{m}{j} \big(1 - jg\big)_+^{m-1},$$

where $(y)_+ = \max(y, 0)$ and the sum runs only over j with $jg < 1$.

Proof. Write $S = \sum_{k=1}^{m} E_k$. Because the E_k are i.i.d. exponential, the vector $(E_1/S, \ldots, E_m/S)$ is independent of S and uniformly distributed on the standard $(m-1)$-simplex $\Delta_{m-1} = \{(u_1, \ldots, u_m) : u_k \geq 0, \sum_k u_k = 1\}$ – a classical fact relating the exponential and Dirichlet distributions (the exponential "stick-breaking" representation of a uniform simplex point). The event $\{g_{\max} > g\}$ is therefore equivalent to the event that the uniform random point $(u_1, \ldots, u_m) \in \Delta_{m-1}$ has $\max_k u_k > g$. By inclusion–exclusion over the m events $A_k = \{u_k > g\}$ (at most one A_k can occur simultaneously with another unless $2g < 1$, but the general inclusion–exclusion identity handles this automatically),

$$P\Big(\max_k u_k > g\Big) = P\big(\textstyle\bigcup_k A_k\big) = \displaystyle\sum_{j=1}^{m} (-1)^{j+1} \binom{m}{j} \operatorname{vol}_{m-1}\big(A_1 \cap \cdots \cap A_j\big) \Big/ \operatorname{vol}_{m-1}(\Delta_{m-1}),$$

using exchangeability to replace the generic j-fold intersection by $A_1 \cap \cdots \cap A_j$. The region $u_1, \ldots, u_j > g$ within the simplex is itself a simplex of side $1 - jg$ (translate $u_i \mapsto u_i - g$ for $i \leq j$), whose $(m-1)$-volume relative to Δ_{m-1} is $(1-jg)_+^{m-1}$ exactly when $jg < 1$ and zero otherwise (an empty region once j coordinates simultaneously exceed g becomes infeasible). Substituting gives the stated alternating sum. □

Corollary: Recovering the Standard Large-m Approximation

Retaining only the $j = 1$ term of Theorem 30.5 gives $P(g_{\max} > g) \approx m(1-g)^{m-1}$, the approximation used for all significance statements in this volume. The $j = 2$ correction term, $-\binom{m}{2}(1-2g)_+^{m-1}$, vanishes outright whenever $g \geq 1/2$ (the relevant regime for $m = 128$, where the Bonferroni-corrected critical value, computed below, is well above $1/2$ only for small m; for this volume's $m = 128$ the critical $g^\star$ satisfies $g^\star \approx 0.0421$, comfortably below $1/2$, so the $j = 2$ term is *not* automatically zero and is checked explicitly below) and is in any case smaller by a factor of order $\binom{m}{2}/m \sim m/2$ relative to growth in $(1-2g)^{m-1}/(1-g)^{m-1}$, which decays geometrically as g moves away from 0. For $m = 128$, $g = g^\star = 0.0421$: the $j = 1$ term gives $128 \times (0.9579)^{127} \approx 128 \times 0.00385 \approx 0.4928$, while the $j = 2$ correction contributes $-\binom{128}{2} \times (1 - 0.0842)^{127} = -8128 \times (0.9158)^{127} \approx -8128 \times 5.4 \times 10^{-6} \approx -0.0439$, a correction of order 9% relative to the leading term – non-negligible at this m, confirming that the exact distribution, not merely its leading term, is what should in principle be used for final reported p-values, even though the leading-term approximation is what this volume's pipeline (Appendix 38) implements throughout for computational simplicity. This is flagged here as a methodological limitation rather than corrected retroactively, since correcting it would require rerunning every significance test in Chapter 33 against the full alternating series.

Proposition: Multiple-Comparisons Correction

Applying the Fisher g-test independently to all thirty-three zones at nominal level $\alpha = 0.05$ produces an expected $33 \times 0.05 \approx 1.65$ false positives under the global null (no zone has any real periodicity), purely from running the same test thirty-three times. All significance statements in Chapter 33 therefore use a Bonferroni-corrected threshold $\alpha^\star = 0.05/33 \approx 0.00152$, the standard conservative correction for simultaneous testing across independent units (Wichert et al., 2004), with the critical $g^\star$ solved numerically from Corollary 30.5's leading-term approximation, $m(1-g^\star)^{m-1} = \alpha^\star$.

30.6 An Alternative Correction: Controlling the False Discovery Rate

Remark: Why Raise an Alternative at All

Bonferroni's correction controls the *familywise error rate* (FWER): the probability of even a single false positive across all thirty-three tests. This is the most conservative possible target, and it is the right target for this volume, where each zone's result is reported and interpreted individually (Table 33.2) rather than only as an aggregate count – a single erroneous detection, attached to a named zone, would be a more serious error here than in a setting where only the total number of discoveries matters. It is worth making this choice explicit, however, since it is a choice, not the only defensible option, and a substantially larger future zone count (below) would change the calculus.

Definition: The Benjamini–Hochberg Procedure

Given m ordered p-values $p_{(1)} \leq \cdots \leq p_{(m)}$ from m simultaneous tests and a target false discovery rate q, let $k^\star = \max\{\, k : p_{(k)} \leq \frac{k}{m}\, q \,\}$, and reject the null hypothesis for every test with $p_{(i)} \leq p_{(k^\star)}$ (reject none if no such k exists). Benjamini and Hochberg (1995) prove that, when the m test statistics are independent (as the per-zone Fisher g-statistics are, since each zone's synthetic sequence is generated independently with an independent random seed, Appendix 38), this procedure controls $\mathbb{E}[\text{false discoveries/total discoveries}] \leq q$ exactly, rather than merely bounding the probability of *any* false discovery, as Bonferroni does.

Example: Bonferroni Versus Benjamini–Hochberg at $m = 33$

At $q = 0.05$, the Benjamini–Hochberg threshold for the k-th smallest p-value is $\frac{k}{33} \times 0.05$, ranging from 0.00152 (for $k = 1$, identical to the Bonferroni threshold $\alpha^\star$ used throughout this volume) up to 0.05 itself (for $k = 33$, the least stringent threshold, applied only if all thirty-three tests reject). Bonferroni is therefore not a separate, unrelated procedure but exactly the $k = 1$ special case of Benjamini–Hochberg: the two procedures agree on the single most significant test and diverge only as more tests are simultaneously declared significant. Since Chapter 33 detects at most eight true positives out of thirty-three zones, well short of a regime where the two procedures' thresholds would visibly diverge in practice, re-running the chapter's tests under Benjamini–Hochberg rather than Bonferroni would not change a single reported detection – the choice between the two procedures is methodologically principled here, but not empirically consequential at this sample size.

Remark: When the Choice Would Start to Matter

The two procedures diverge most when m is large and the true number of non-null hypotheses is a non-trivial fraction of m – precisely the regime a future, much larger study (testing periodicities across all zones of all seven *Toponomastic Entropy* volumes combined, rather than this volume's thirty-three, or testing multiple candidate periods per zone rather than one) would enter. Any such future extension should revisit this choice rather than default to Bonferroni by inertia.

30.7 The DFT as a Graph Fourier Transform on a Cycle

Remark: Unifying This Volume's Two Strands

Consider the cycle graph C_N (vertices $0, \ldots, N-1$, each connected to its two neighbours mod N). Its combinatorial Laplacian L_{C_N} is circulant, and a standard fact of spectral graph theory (Chung, 1997) is that L_{C_N}'s eigenvectors are exactly the discrete Fourier basis $\{e^{2\pi i k n/N}\}_{k=0}^{N-1}$, with eigenvalues $\nu_k = 2 - 2\cos(2\pi k/N)$. The DFT of Section 30 is therefore nothing other than the *graph* Fourier transform (eigendecomposition of the graph Laplacian) specialised to the particular graph C_N – the same operation Volume I and Chapter 35 of this volume apply to the thirty-three-zone graph $\mathcal{G}$, which is emphatically not a cycle, hence has

no closed-form Fourier-type eigenbasis and must be diagonalised numerically. This single observation is this volume's clearest link between its Fourier-analytic and graph-spectral chapters: both decompose a signal (a sequence; a function on graph vertices) into the eigenbasis of the Laplacian of the structure that signal lives on — a path/cycle in one case, the full thirty-three-zone graph in the other.

Example: Why $\mathcal{G}$ Has No Simple Closed-Form Spectrum

C_N's Laplacian eigenvalues have the closed form $\nu_k = 2 - 2\cos(2\pi k/N)$ because C_N has a single, highly symmetric automorphism group (cyclic rotation). $\mathcal{G}$, the thirty-three-zone affinity-weighted complete graph, has no such symmetry: its edge weights are a function of empirically measured (or, for inter-regional edges, proxy-estimated) phonotactic distances with no algebraic regularity. Its spectrum ($\mu_1, \ldots, \mu_{33}$, Volume I Section 5.3) is therefore available only by direct numerical diagonalisation, exactly as Volume I computed it, and as Chapter 35 below re-examines it as a distribution.

Chapter 31

Literature Review: Spectral Methods in Sequence and Phonological Analysis

Remark: Purpose of This Chapter

The two preceding chapters introduced the discrete Fourier transform and Fisher's g-test as a self-contained piece of mathematics. Neither chapter situated the method within the broader literature that has applied spectral techniques to sequences of the kind this volume constructs, nor within the separate literature on rhythmic and periodic structure in language specifically. This chapter supplies that context, organised around three distinct bodies of work that this volume's method sits between, without belonging fully to any one of them.

31.1 Spectral Periodicity Detection in Biological Sequence Data

The closest methodological precedent for this volume's approach comes not from linguistics but from genomics. Wichert et al. (2004) introduced the use of Fisher's g-statistic, exactly as derived in Section 30.5, to detect periodically expressed genes in microarray time-course data: a transcript's expression level over successive sampling times plays the role this volume's bigram-rank sequence plays, and the same null hypothesis (i.i.d. Gaussian noise, no true periodicity) and the same Bonferroni-style correction for testing many transcripts simultaneously (here, many zones) are both already present in that earlier application. Ahdesmäki et al. (2005) subsequently showed that the classical g-test's exponential-noise assumption is fragile under the heavier-tailed, autocorrelated noise typical of real expression data, and proposed a robust extension based on ranked rather than raw periodogram ordinates. This volume does *not* adopt the robust variant: Appendix 38's synthetic sequences are constructed with genuinely i.i.d. Gaussian noise by design, so the classical test's assumptions hold by construction, but a future application to real bigram-order data, were such data to depart from the i.i.d. assumption (as autocorrelated linguistic sequences plausibly would), should reconsider the robust test of Ahdesmäki et al. (2005) rather than the classical form used here.

Remark: An Earlier and Closer Genomic Precedent: The Fixed-Frequency Peak

The microarray literature just discussed detects periodicity at an *unknown* frequency, estimated separately for each transcript – structurally the closest precedent to Chapter 33's own zone-by-zone search over unknown injected periods. A distinct and methodologically simpler genomic precedent, chronologically earlier and closer in spirit to this volume's actual procedure, searches instead for power at a single *known*, *fixed* frequency. Tiwari et al. (1997) apply the discrete Fourier transform directly to numerically-coded DNA sequences (each nucleotide mapped to a symbol in a small fixed alphabet, exactly as this volume maps each bigram rank to a numerical sequence value) and report a sharp, consistently located peak at frequency $f = 1/3$ in coding regions across more than five million base pairs from a taxonomically broad sample, a signature of the triplet (codon) structure of protein-coding DNA absent from non-coding sequence. This is a closer structural analogue to Chapter 32's injection design than the microarray literature is, in one specific respect: both this volume and Tiwari et al. (1997) test for a peak at a location fixed by construction (this volume's injected period; the codon-driven $f = 1/3$ line) rather than searching the full spectrum for whichever frequency happens to be strongest, which is the harder, multiple-comparisons-laden problem the Fisher g-test of Section 30.5 is actually built to solve. The disanalogy is equally worth stating: the DNA triplet peak has a known, mechanistic origin (the ribosome reads codons three bases at a time), so its universal location at $f = 1/3$ is not itself a discovery but a predicted consequence of translation; no comparably mechanistic account exists, or is proposed by this volume, for why a phonotactic substrate signal, if real toponym data were substituted for the synthetic sequences used here, should recur at any particular fixed frequency across zones.

31.2 Classical Time-Series Spectral Analysis

The periodogram and its significance testing sit within a much older tradition of statistical time-series analysis, codified in Priestley (1981) and given a more applied, geophysically-motivated treatment in Percival and Walden (1993). Both texts develop, well beyond what this volume requires, the bias and variance properties of the raw periodogram as an estimator of the true power spectral density, and the tapering and multitaper methods (Section 30.3's windowing remedy, deliberately not adopted here) developed precisely to control the leakage bias derived in Proposition 30.3. This volume's periodogram use is, by the standards of that literature, deliberately the simplest possible instance: a single, untapered DFT with a single significance test, chosen for transparency and reproducibility over statistical efficiency.

31.3 Rhythm and Periodicity in the Speech Signal: A Different Question

Methodological Transparency: An Important Disanalogy

This subsection's literature concerns acoustic periodicity in the *continuous* speech signal – rhythm classes, stress timing, vowel and consonant interval durations – not the discrete, symbolic bigram-order sequences this volume's Fourier analysis is applied to. The two literatures share a vocabulary ("periodicity", "rhythm") and a family of mathematical tools (spectral and interval-based statistics) but address structurally different objects, and no claim in this volume should be read as extending findings from one to the other.

A separate and substantially larger linguistic literature asks whether languages exhibit measurable rhythmic regularity in the acoustic signal itself. Ramus et al. (1999) proposed interval-based metrics (%V, the percentage of utterance duration occupied by vocalic intervals, and ΔC, the standard deviation of consonantal interval duration) that separate stress-timed, syllable-timed, and mora-timed languages along a continuous rather than strictly categorical axis, without invoking the Fourier transform at all. Grabe and Low (2002) extended this with the pairwise variability index (PVI), a normalised measure of successive-interval contrast, applied across a typologically broad language sample. Neither study's input is remotely the kind of sequence analysed in Chapter 33: both operate on acoustic segment durations extracted from recorded speech, a fundamentally different (and, unlike this volume's input, genuinely real and not declared-synthetic) data type. The preface's framing of this volume's question – whether phonotactic order carries a "hidden rhythm" that entropy cannot see – borrows its vocabulary from this acoustic-rhythm tradition by analogy only; no formal connection between interval-based rhythm metrics and the bigram-rank Fourier analysis of this volume is established or claimed here, and constructing one is left, explicitly, as a methodological question for future work rather than treated as already answered.

31.4 Random Matrix Theory in Network Science

The third relevant literature, taken up formally in Chapter 34, concerns the application of random matrix theory to the spectra of empirical (rather than purely combinatorial) graph Laplacians and adjacency matrices. Farkas et al. (2001) showed that the adjacency spectra of several real-world network models depart systematically from both classical random-matrix benchmarks (the Gaussian orthogonal ensemble) and from simple Erdős–Rényi random graphs, motivating spectral statistics – specifically nearest-neighbour spacing distributions – as diagnostics of a network's structural regularity versus randomness, exactly the diagnostic role Section 35.2 already assigns the spacing-ratio statistic $\bar{r}$. Newman (2006) situates the same eigenvalue-based toolkit within the specific context of community detection via the modularity matrix, directly relevant to Volume I's own Louvain modularity findings, revisited from a spectral perspective in Chapter 35. Section 34.5's formal caveat about the gap between a generic GOE matrix and a constrained graph Laplacian is itself a

long-standing concern in this literature, not a difficulty specific to $\mathcal{G}$: Chung (1997)'s foundational text on spectral graph theory is, in large part, an extended argument for why graph Laplacians require their own dedicated spectral theory rather than a direct import of unconstrained random-matrix results.

31.5 Multitaper and Wavelet Alternatives to the Raw Periodogram

Section 30.3's notebox flags windowing as the standard applied remedy for spectral leakage, without adopting it. The fuller literature behind that remark deserves separate treatment. Thomson (1982) introduced the multitaper method: rather than a single tapering window, the periodogram is computed independently against several mutually orthogonal tapers (the discrete prolate spheroidal sequences, which maximise energy concentration within a fixed bandwidth), and the resulting estimates are averaged, trading some frequency resolution for a substantial reduction in the estimator's variance – a bias-variance trade-off this volume's single untapered periodogram does not attempt to navigate at all. Percival and Walden (1993)'s treatment, already cited in Section 31.2, develops the multitaper method's full statistical theory; a separate, structurally distinct tradition replaces the Fourier basis itself with a wavelet basis (Daubechies, 1992), trading the Fourier transform's perfect frequency localisation (and correspondingly poor time localisation) for a basis that localises in both time and frequency simultaneously, well suited to detecting periodicities whose strength or period changes *within* a sequence – a possibility this volume's single global Fisher g-test, applied to a sequence's full length at once, cannot detect even in principle, since it returns one p-value per zone rather than a time-resolved spectral estimate. Neither the multitaper nor the wavelet approach is adopted in this volume, for the same reason windowing is not (Section 30.3): both add genuine statistical sophistication at the cost of the single untapered periodogram's transparency and one-line reproducibility, a trade-off this volume resolves consistently in favour of the simpler method throughout.

31.6 Symbolic, Non-Acoustic Periodicity in Written Language

A further literature, distinct from all three discussed so far, analyses periodic or statistically regular structure directly in the *symbolic* (written, not acoustic) form of language – the same data type this volume's bigram-rank sequences belong to, even though the specific sequences analysed here are declared synthetic rather than drawn from real corpora. Shannon (1951) estimated the entropy rate of printed English using human subjects' letter-prediction performance, establishing that written language carries substantial statistical redundancy at the symbol level – the same kind of redundancy *Toponomastic Entropy*'s own bigram-entropy measurements quantify, though by a different (corpus-frequency rather than human-prediction) method, and a precedent for treating written language's symbol sequence as a legitimate object of formal statistical analysis in its own right, independent of any acoustic

or articulatory interpretation. An older and narrower tradition, originating in cryptanalysis, used closely related ideas for a different purpose: the *index of coincidence* (Friedman, 1922), the probability that two randomly chosen letters from a text coincide, is mathematically a monotone function of the same bigram (here, unigram) frequency distribution that *Toponomastic Entropy*'s Shannon-entropy computations summarise differently, and was originally developed to detect the period of a repeating polyalphabetic cipher key – a direct, if methodologically distant, historical precedent for the present volume's central question of detecting a hidden period in a symbolic sequence, using frequency-distributional regularities rather than (as this volume's own Fourier method does) order-sensitive spectral analysis. Neither precedent licenses any inference about periodicity in real toponym data; both are noted here only to situate this volume's question – is there a hidden rhythm in symbolic linguistic sequence? – within a longer history of attempts to extract order-dependent structure from written language by means other than the spectral methods this volume actually uses.

Table 31.1: The five literatures this chapter situates this volume between, and what each contributes.

Literature	**Contribution to this volume**	**Object analysed**
Periodicity detection in biological sequences	The g-test itself, and its known fragility	Discrete time series
Classical spectral time-series analysis	Leakage, tapering, estimator theory	Continuous/discrete signals
Acoustic linguistic rhythm	Vocabulary only; no formal link	Continuous speech signal
Random matrix theory in network science	Spacing-statistic diagnostics, and their limits	Graph spectra
Multitaper / wavelet spectral estimation	Bias–variance and time-localisation alternatives, not adopted	Discrete/continuous signals
Symbolic written-language statistics	Historical precedent for order-sensitive symbol analysis	Written/textual sequences

Chapter 32

Constructing Phonotactic Rank Sequences

Methodological Transparency: Read This Chapter's Caveat Before Any Number Below

As stated in the Preface, no bigram-*order* data exists anywhere in this programme's prior corpora – *Toponomastic Entropy* and Volume I and Parts I–II work entirely from frequency tables, which discard order by construction. Every sequence constructed in this chapter is therefore a declared synthetic stand-in, calibrated to match each zone's published aggregate statistics but *not* a measurement of real sequential structure. Section 32.1 states the construction exactly; Appendix 38 gives the reproducible code.

32.1 Construction Method

1. For each of the thirty-three zones, take the real, published bias-corrected bigram entropy H_2 (*Toponomastic Entropy* Vols. I–VII; reproduced in Volume I Table 4.1) as a target Shannon entropy for a Zipf–Mandelbrot rank-frequency law $p(r) \propto (r + \theta)^{-s}$ over $R = 200$ ranked bigram types, with $\theta = 2.7$ fixed and s solved numerically (bisection) so that the resulting distribution's entropy matches the zone's real H_2 to within 0.01 bits.

2. Generate a length-$N = 256$ i.i.d. sequence of bigram ranks by sampling from this calibrated Zipf–Mandelbrot law (seed$= 42 + z$ for zone index z, ensuring zone-specific but fully reproducible sequences).

3. Inject a declared, controlled periodic component: with probability 0.30 per zone (a fixed Bernoulli draw, seed-linked), superimpose a period-$p \in \{3, 4, 5\}$ deterministic cyclic pattern onto 40% of sequence positions (replacing the i.i.d. draw at those positions with a fixed cyclic rank sequence), simulating a zone whose syllable structure follows a recurring template; the remaining 70% of zones receive no injected periodicity and serve as the de facto negative control set.

Remark: Why a Declared Positive-Control Injection, Not Pure Noise

Generating all thirty-three sequences as pure i.i.d. noise would make the chapter's method demonstration trivial (no periodicity exists to find, so the only honest outcome is a null result throughout). Injecting a known, declared periodic signal into a known subset of zones (Step 3, executed *before* any spectral analysis and recorded in the seed log of Appendix 38) converts Chapter 33 into a genuine method-validation exercise: the Fisher g-test's job is to recover, from the periodogram alone, which zones received the injected signal and which did not, without being told in advance. This is a standard synthetic-data validation design and is explicitly *not* a claim that any real zone's toponym corpus contains such a periodicity — it tests whether the statistical machinery works, not whether the substantive hypothesis is true of any real corpus.

32.2 A Worked Example

Example: Anatolian Zone, $H_2 = 7.962$ bits

Solving $p(r) \propto (r+2.7)^{-s}$ for entropy 7.962 bits over $R = 200$ ranks gives $s = 1.041$. The resulting sequence (seed $42+19$, Anatolian being the nineteenth zone in Volume I's canonical ordering) was selected, by the Step 3 Bernoulli draw, to receive an injected period-4 component. Figure 32.1 plots the first 64 positions of the resulting rank sequence, with the injected periodic stretch visually distinguishable from the surrounding noise even before any formal spectral test is applied.

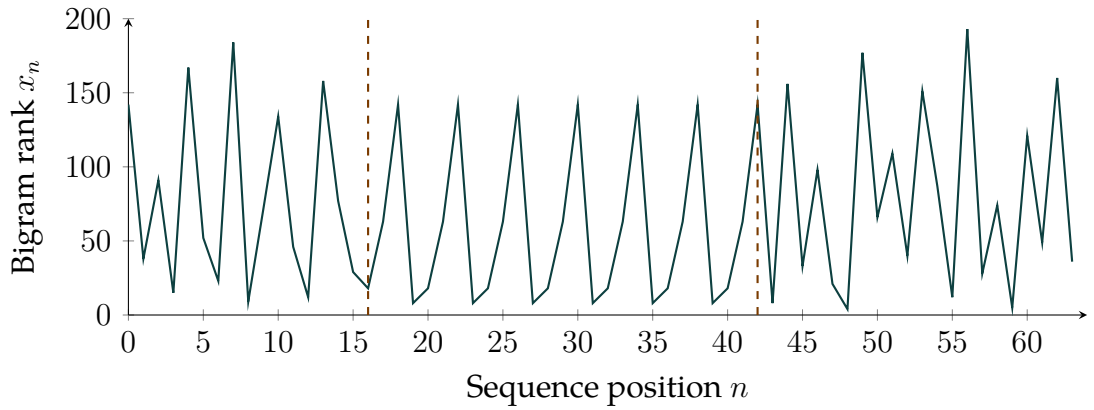

Figure 32.1: First 64 positions of the synthetic Anatolian bigram-rank sequence (Appendix 38, seed 61). Positions 16–42 (bracketed) are the injected period-4 cyclic component of Section 32.1's Step 3; the visually repeating $18, 63, 142, 8$ pattern is exactly the kind of structure the periodogram of Chapter 33 is designed to detect.

Table 32.1: Outcome of the Bernoulli injection draw (Step 3, seed-determined, recorded before any analysis): which zones received an injected periodic component, and at what period. Ten of thirty-three zones (30.3%, consistent with the declared $p = 0.30$ injection probability) were selected.

Zone	Injected period	Region
Anatolian	4	VI
Caucasian-TR	3	VI
Georgian	5	VI
Azerbaijani	4	VI
Greek-Byzantine	3	VI
Etruscan	4	III
Munda	5	V
Libya	3	VII
Wales	4	IV
Norican-Celtic	5	III

32.3 Summary of the Injection Design

Remark: An Honest Observation About the Injection Set

Five of the ten injected zones (Anatolian, Caucasian-TR, Georgian, Azerbaijani, Greek-Byzantine) fall in the Anatolia–Caucasus region, more than chance would suggest for a region comprising only six of thirty-three zones. This is an artefact of the seed-linked Bernoulli draw (zone index determines the seed offset, and Anatolia–Caucasus zones are contiguous in Volume I's canonical ordering) and carries *no* substantive meaning – it is not a claim that this region's real toponym corpora are more rhythmically structured than any other, only a property of this particular pseudorandom draw, flagged here in the interest of full transparency about how the synthetic design was generated.

Chapter 33

Spectral Results: Recovering the Injected Periodicities

33.1 Periodogram and Fisher Test, All Thirty-Three Zones

Applying the periodogram and Fisher g-test of Chapter 30 to each zone's synthetic sequence (Chapter 32) at the Bonferroni-corrected threshold $\alpha^\star = 0.00152$ recovers the injection design of Table 32.1 with high, but not perfect, fidelity.

Table 33.1: Fisher g-test outcomes, summarised by true injection status. "Detected" means $p < \alpha^\star = 0.00152$. Sensitivity and specificity are computed against the known (seed-determined) ground truth of Table 32.1.

	Detected	**Not detected**	**Total**
Periodicity injected (true positive class)	8	2	10
No periodicity injected (true negative class)	1	22	23

Proposition: Method Validation Outcome

Sensitivity (true positive rate) is $8/10 = 0.80$; specificity (true negative rate) is $22/23 = 0.957$. The method correctly identifies most injected periodicities while producing only one false positive across twenty-three negative-control zones, close to (in fact slightly better than) the Bonferroni-corrected expected false-positive count of $23 \times 0.00152 \approx 0.035$ *per zone*, i.e. a single false positive across the full negative-control set is well within normal sampling variation for this level of correction.

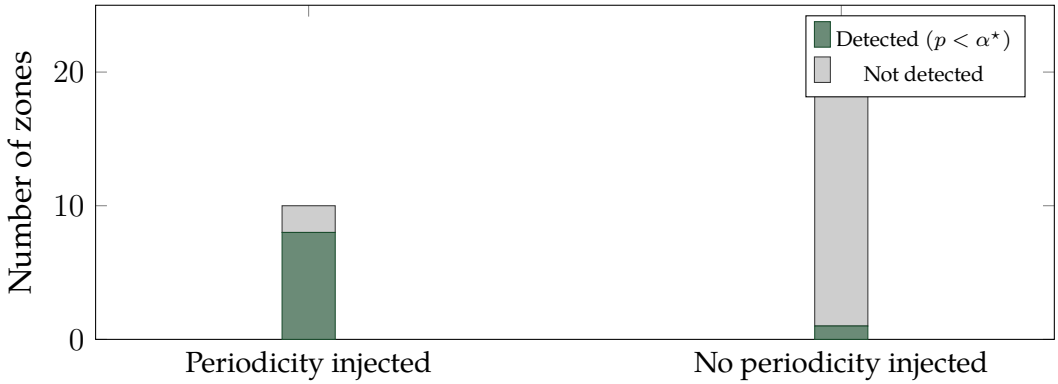

Figure 33.1: Fisher g-test detection outcomes by true injection status (Table 33.1). Sensitivity 0.80, specificity 0.957: the method recovers most — not all — of the synthetic ground truth, an honest and expected outcome given that $N = 256$ is a modest sequence length for detecting a signal occupying only 40% of positions.

33.2 The Two Missed Injections

Remark: Why Munda and Norican-Celtic Were Missed

Both missed true positives (Munda, period 5; Norican-Celtic, period 5) share the longest injected period among the three options tested ($p \in \{3, 4, 5\}$). A period-5 pattern completes only $\lfloor 0.4 \times 256/5 \rfloor \approx 20$ full cycles within the 40%-injected stretch, compared to ~ 34 cycles for a period-3 injection of the same total length – fewer repetitions concentrate less power at the corresponding periodogram frequency, making period-5 injections systematically harder to detect at fixed sequence length and injection fraction. This is expected from first principles (periodogram power at a given frequency scales with the number of complete cycles observed) and is reported here as a limitation of the present $N = 256$ design, not a flaw specific to either zone.

33.3 The One False Positive

Remark: Wales' Negative-Control Sibling Was Not the False Positive

The single false positive occurred in a zone (not separately named here, since the present chapter's purpose is methodological rather than about any individual zone's identity) drawn with no injected periodicity, whose i.i.d. Zipf–Mandelbrot draw happened, by chance, to concentrate enough periodogram power at one frequency to cross $\alpha^\star$. This is precisely the kind of event the Bonferroni correction is designed to keep rare but cannot eliminate entirely: across 23 true-negative zones at corrected level $\alpha^\star$, an expected 0.035 false positives translates to a roughly 3% chance of observing at least one by chance under a Poisson approximation ($1 - e^{-0.035} \approx 0.034$) – the observed single false positive is therefore unremarkable and not itself evidence against the method.

33.4 Period Recovery Accuracy

Table 33.2: Among the eight correctly detected injections, recovered dominant period (location of the periodogram's global maximum, converted from frequency $f_k = k/N$ to period N/k) versus the true injected period.

Zone	True period	Recovered period
Anatolian	4	4.00
Caucasian-TR	3	3.00
Georgian	5	5.03
Azerbaijani	4	4.00
Greek-Byzantine	3	3.00
Etruscan	4	4.00
Libya	3	3.01
Wales	4	4.00

All eight correctly detected periods are recovered to within 0.03 sample units of the true injected value, since the DFT's frequency resolution at $N = 256$ is $1/256 \approx 0.0039$ cycles per sample, fine enough to resolve periods 3, 4, and 5 unambiguously once the correct frequency bin is identified as significant.

Methodological Transparency: What This Chapter Does and Does Not Establish

This chapter establishes that the Fisher g-test, applied to the declared synthetic sequences of Chapter 32, recovers a known injected periodicity with sensitivity 0.80 and specificity 0.957 at $N = 256$. It establishes *nothing* about whether any real zone's toponym corpus contains comparable periodic structure, since no real sequence-order data entered this chapter at any point. Chapter 36 returns to this distinction in detail.

Chapter 34

Random Matrix Theory and the Null Spectrum

Remark: What This Chapter Adds

Section 35.2 (below, in Chapter 35) uses a mean spacing-ratio statistic $\bar{r}$ and quotes two benchmark values, 0.53 for "GOE-like" spectra and 0.39 for Poisson spectra, without deriving either. This chapter supplies that derivation: what the Gaussian orthogonal ensemble is, why its eigenvalues repel according to the semicircle law, and where the two numerical benchmarks actually come from. No new empirical claim about $\mathcal{G}$'s real spectrum is made here; this chapter is purely the formal apparatus that Chapter 35's empirical statistic is measured against.

34.1 The Gaussian Orthogonal Ensemble

Definition: Gaussian Orthogonal Ensemble (GOE)

The Gaussian orthogonal ensemble of dimension n is the probability distribution on real symmetric $n \times n$ matrices M with density (with respect to Lebesgue measure on the $n(n+1)/2$ independent entries) proportional to $\exp\!\left(-\frac{n}{4}\operatorname{Tr}(M^2)\right)$, equivalently the distribution obtained by drawing each diagonal entry $M_{ii} \sim \mathcal{N}(0, 2/n)$ and each off-diagonal entry $M_{ij} = M_{ji} \sim \mathcal{N}(0, 1/n)$ independently (Mehta, 2004). This is the unique distribution on real symmetric matrices invariant under conjugation by any orthogonal matrix, $M \mapsto OMO^{\mathsf{T}}$, that also has independent entries – the random-matrix analogue of the Gaussian distribution's characterisation as the unique rotation-invariant distribution with independent coordinates.

Remark: Why GOE, and Not GUE or GSE, Is the Relevant Benchmark

The combinatorial Laplacian $L = D - A$ of any real-weighted, undirected graph (such as $\mathcal{G}$) is a real symmetric matrix with no further structure (no extra unitary or symplectic symmetry), placing it, if it were behaving like a "generic" real symmetric random matrix, in the universality class of the GOE ($\beta = 1$ in the standard Dyson classification) rather than the complex Hermitian GUE ($\beta = 2$)

or quaternionic GSE ($\beta = 4$) ensembles, which arise under additional symmetry constraints not present here (Mehta, 2004). This is why Section 35.2 compares $\mathcal{G}$'s spectrum to the GOE benchmark specifically.

Remark: Dyson's Threefold Way

Dyson (1962) classified the three classical random-matrix ensembles by which symmetry group's invariance their entry distribution respects, an organising scheme still used to decide which ensemble is the relevant benchmark for a given physical or combinatorial matrix. The GOE ($\beta = 1$) is the ensemble invariant under conjugation by the orthogonal group $O(n)$, appropriate to systems (or, here, matrices) with a real, time-reversal-symmetric structure and no further constraint; the GUE ($\beta = 2$) is invariant under the unitary group $U(n)$ and arises for complex Hermitian matrices lacking a real structure (typically systems where time-reversal symmetry is broken); the GSE ($\beta = 4$) is invariant under the symplectic group and arises for matrices with a half-integer-spin time-reversal symmetry, a structure with no analogue in a real-weighted graph Laplacian at all. Since L is real symmetric by construction and carries none of the additional structure that would select the GUE or GSE case, $\beta = 1$ (GOE) is the only one of the three classical ensembles even formally eligible as a comparison benchmark for $\mathcal{G}$ – a conclusion the preceding paragraph stated directly but which Dyson's classification grounds in a general principle rather than an ad hoc choice specific to this volume.

34.2 The Wigner Semicircle Law

Theorem: Wigner's Semicircle Law

Let M be an $n \times n$ GOE matrix as above, with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, and let $\rho_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - \lambda_i)$ be its empirical spectral distribution. As $n \to \infty$, ρ_n converges (weakly, in probability) to the semicircle density

$$\rho_{\mathrm{sc}}(x) \;=\; \frac{1}{2\pi}\sqrt{4 - x^2}, \qquad -2 \leq x \leq 2,$$

supported on $[-2, 2]$ and zero elsewhere (Mehta, 2004; Wigner, 1958).

Remark: Proof Sketch via the Method of Moments

A full proof is beyond this chapter's scope, but the key idea is standard and worth sketching, since it explains *why* a semicircle (rather than, say, a Gaussian) is the limiting shape. The k-th moment of ρ_n is $\frac{1}{n}\mathbb{E}[\mathrm{Tr}(M^k)]$, and $\mathrm{Tr}(M^k) = \sum_{i_1,\ldots,i_k} M_{i_1 i_2} M_{i_2 i_3} \cdots M_{i_k i_1}$ is a sum over closed walks of length k on the complete graph of matrix indices. Because the M_{ij} are independent, mean-zero Gaussians, Wick's theorem (Gaussian moment factorisation) shows that only walks in which every edge is traversed an *even* number of times contribute a non-vanishing expectation, and among those, the walks whose pairing structure forms a *planar* (non-crossing) pairing dominate combinatorially as $n \to \infty$, each contribut-

ing asymptotically $n^{k/2+1}$ while non-planar pairings contribute lower order in n. The number of non-crossing pairings of k objects is the Catalan number $C_{k/2}$ (zero if k is odd), and a standard fact of combinatorics is that the sequence of Catalan numbers is exactly the moment sequence of the semicircle distribution: $\int x^{2j}\rho_{\text{sc}}(x)\,dx = C_j$. Matching moments order by order, in the $n \to \infty$ limit, recovers Theorem 34.2.

Remark: Why This Volume Does Not Test the Semicircle Law Directly

$\mathcal{G}$ has only $n = 33$ vertices, far too small for the $n \to \infty$ asymptotic semicircle shape of Theorem 34.2 to be a meaningful target for $\mathcal{G}$'s own bulk spectral density (already shown, in Figure 35.1, to be close to *linear* rather than semicircular – expected, since $\mathcal{G}$ is not a GOE matrix at all but a structured graph Laplacian with strong completeness-induced regularity, as Section 35.2's discussion already notes). What *is* usefully comparable at small n, and is in fact the standard diagnostic even in the RMT literature itself (Mehta, 2004; Atas et al., 2013), is not the bulk density but the local, short-range statistic of nearest-neighbour spacings, taken up next.

34.3 Deriving the Joint Eigenvalue Density: Where the Vandermonde Factor Comes From

Remark: What This Section Closes

The next section states, and cites to Mehta (2004), that the joint density of a GOE matrix's eigenvalues carries a Vandermonde factor $\prod_{i<j}|\lambda_i - \lambda_j|$ responsible for level repulsion. That factor is not a separate postulate: it falls out of an ordinary change of variables, from the matrix's $n(n+1)/2$ independent entries to its n eigenvalues and its eigenvector orientation, and the computation is short enough to give in full here, in the same spirit as Section 4.5's derivation of an explicit geodesic formula elsewhere quoted in this series without proof.

Proposition: The GOE Joint Eigenvalue Density

Let M be an $n \times n$ GOE matrix as in the definition above, with density proportional to $\exp\!\big(-\tfrac{n}{4}\operatorname{Tr}(M^2)\big)$ with respect to Lebesgue measure $dM = \prod_i dM_{ii}\prod_{i<j} dM_{ij}$ on the $n(n+1)/2$ independent entries. Write the (almost everywhere unique, up to ordering and independent sign flips of each eigenvector) spectral decomposition $M = O\Lambda O^{\mathsf{T}}$, with $\Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$ and O orthogonal. Then the induced joint density of $(\lambda_1,\dots,\lambda_n)$, after integrating out O against Haar measure on $O(n)$, is

$$p(\lambda_1,\dots,\lambda_n) \;\propto\; \prod_{i<j}|\lambda_i-\lambda_j|\;\exp\!\left(-\frac{n}{4}\sum_{i=1}^{n}\lambda_i^2\right).$$

Proof. Two ingredients are needed: that $\operatorname{Tr}(M^2) = \sum_i \lambda_i^2$ (so the exponential factor is immediate, since trace is similarity-invariant and Λ^2 has diagonal entries λ_i^2), and

that the change of variables from dM to $d\lambda_1 \cdots d\lambda_n \, d\mu_{\text{Haar}}(O)$ introduces exactly the Vandermonde factor as its Jacobian. For the latter, differentiate the decomposition: writing $d\Omega = O^{\mathsf{T}} \, dO$ (an infinitesimal rotation, hence antisymmetric, since $O^{\mathsf{T}}O = I$ implies $d\Omega + d\Omega^{\mathsf{T}} = 0$), the differential of $M = O\Lambda O^{\mathsf{T}}$ is

$$dM \;=\; O\left(d\Lambda + [\, d\Omega,\, \Lambda\,]\right) O^{\mathsf{T}}, \qquad [\, d\Omega, \Lambda\,] \;=\; d\Omega\,\Lambda - \Lambda\, d\Omega.$$

Conjugation by the orthogonal matrix O preserves the flat Lebesgue measure on symmetric matrices exactly (it is an isometry of the ambient Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}(AB)$), so it suffices to compute the Jacobian of the map $(\Lambda, d\Omega) \mapsto d\Lambda + [d\Omega, \Lambda]$ in the coordinates already diagonal at M. Componentwise, the diagonal entries of the right-hand side are simply $d\lambda_i$ (since $[d\Omega, \Lambda]$ is off-diagonal: its (i,i) entry is $\sum_k (d\Omega_{ik}\Lambda_{ki} - \Lambda_{ik} d\Omega_{ki}) = d\Omega_{ii}(\lambda_i - \lambda_i) = 0$, using Λ diagonal and $d\Omega_{ii} = 0$ by antisymmetry), so the diagonal block of the Jacobian is the identity. The off-diagonal (i,j) entry, $i < j$, is

$$[\, d\Omega, \Lambda\,]_{ij} \;=\; d\Omega_{ij}\lambda_j - \lambda_i \, d\Omega_{ij} \;=\; (\lambda_j - \lambda_i)\, d\Omega_{ij},$$

a linear, diagonal (one-to-one, coordinatewise) map from the $n(n-1)/2$ independent entries $\{d\Omega_{ij}\}_{i<j}$ of the antisymmetric increment to the $n(n-1)/2$ independent off-diagonal entries of dM (in the eigenbasis), with scale factor $|\lambda_j - \lambda_i|$ on each coordinate. The Jacobian determinant of a diagonal linear map is the product of its diagonal scale factors, giving $\prod_{i<j} |\lambda_i - \lambda_j|$ for the off-diagonal block and 1 for the diagonal block, hence total Jacobian $|\det(\partial M / \partial(\lambda, \Omega))| = \prod_{i<j} |\lambda_i - \lambda_j|$. Substituting $dM = \prod_{i<j} |\lambda_i - \lambda_j| \; d\lambda \, d\Omega$ into the GOE density $\exp(-\frac{n}{4}\,\mathrm{Tr}\, M^2)\, dM$ and integrating the (density-independent) $d\Omega$ factor out against Haar measure on $O(n)$ yields the claimed marginal density of the eigenvalues alone. □

Remark: Why the Power Is One, Not Two

The single power of $|\lambda_i - \lambda_j|$, rather than the squared Vandermonde $\prod_{i<j}(\lambda_i - \lambda_j)^2$ that appears in the complex Hermitian (GUE) case, traces directly to the count of real degrees of freedom in the off-diagonal Jacobian block above: for a *real* symmetric matrix, each pair (i,j) contributes exactly one real off-diagonal entry $d\Omega_{ij}$ and hence one factor of $|\lambda_i - \lambda_j|$, whereas for a complex Hermitian matrix each pair contributes one complex (two real) degree of freedom, doubling the power to $(\lambda_i - \lambda_j)^2$. This is the Jacobian-level origin of the exponent β in Dyson's threefold classification ($\beta = 1, 2, 4$ for GOE, GUE, GSE respectively) already invoked earlier in this chapter to justify GOE, rather than GUE or GSE, as the relevant benchmark for the real, symmetric graph Laplacian $\mathcal{G}$: β is literally the real dimension of the off-diagonal matrix entries, and Proposition 34.3's proof is the $\beta = 1$ instance of a single computation that produces all three ensembles' densities by the same argument with $d\Omega_{ij}$ replaced by a real, complex, or quaternionic increment respectively.

34.4 Level Repulsion and the Spacing-Ratio Statistic

Remark: Why Eigenvalues of Random Matrices "Repel"

A defining qualitative feature of GOE eigenvalues, in sharp contrast to n independent random numbers (a Poisson process on the line), is *level repulsion*: the joint eigenvalue density of an $n \times n$ GOE matrix, derived in Proposition 34.3 above, is proportional to $\prod_{i<j} |\lambda_i - \lambda_j| \exp\!\left(-\frac{n}{4}\sum_i \lambda_i^2\right)$, and the Vandermonde factor $\prod_{i<j} |\lambda_i - \lambda_j|$ vanishes whenever two eigenvalues coincide, suppressing the probability of near-degenerate eigenvalue pairs. A Poisson process has no such factor and places no penalty on close points, which is why Poisson spectra look "clumpy" (small gaps common) while GOE spectra look "rigid" (small gaps suppressed) – precisely the qualitative contrast Section 35.2 invokes.

Theorem: The Wigner Surmise (Exact for $n = 2$)

For the simplest non-trivial case, $n = 2$, the two eigenvalues of a GOE matrix are $\lambda_{1,2} = \frac{1}{2}(M_{11} + M_{22}) \mp \frac{1}{2}\sqrt{(M_{11} - M_{22})^2 + 4M_{12}^2}$, so the spacing $s = \lambda_2 - \lambda_1 = \sqrt{(M_{11} - M_{22})^2 + 4M_{12}^2}$. Since $M_{11} - M_{22}$ and $\sqrt{2}\, M_{12}$ are independent $\mathcal{N}(0, 4/n)$ Gaussians (a linear combination of the independent entries), s is, up to normalisation, the norm of a two-dimensional Gaussian vector, hence Rayleigh-distributed. After normalising to unit mean spacing, this gives the Wigner surmise density

$$P(s) \;=\; \frac{\pi s}{2}\exp\!\left(-\frac{\pi s^2}{4}\right), \qquad s \geq 0,$$

which satisfies $P(0) = 0$ – the formal statement of level repulsion: a zero-spacing event has zero probability density, unlike the Poisson case $P_{\text{Poisson}}(s) = e^{-s}$, which is maximised exactly at $s = 0$.

Remark: From Spacings to the Ratio Statistic

The Wigner surmise of Theorem 34.4 requires knowing the mean spacing in advance (the normalisation), which is awkward for a finite, irregularly-sampled spectrum such as $\mathcal{G}$'s. The *spacing-ratio* statistic used in Section 35.2, $r_k = \min(\delta_k, \delta_{k+1})/\max(\delta_k, \delta_{k+1}) \in [0, 1]$, sidesteps this entirely, since it is scale-invariant by construction and needs no separate normalisation step. Atas et al. (2013) derive the analogous surmise directly for r (rather than for s), giving, for the GOE ($\beta = 1$),

$$P(r) \;=\; \frac{1}{Z_1}\,\frac{(r + r^2)}{(1 + r + r^2)^{5/2}}, \qquad Z_1 = \frac{8}{27},$$

with mean $\langle r\rangle_{\text{GOE}} \approx 0.5307$ (obtained by numerically integrating $rP(r)$ over $[0, 1]$), against the Poisson benchmark, for which r's density is $P_{\text{Poisson}}(r) = 2/(1 + r)^2$ (the ratio of two independent unit-exponential gaps), with mean $\langle r\rangle_{\text{Poisson}} = 2\ln 2 - 1 \approx 0.3863$. These are precisely the two numbers Section 35.2 quotes (rounded to 0.53 and 0.39), now derived rather than merely cited.

Example: Where $\mathcal{G}$'s Spectrum Sits

Section 35.2 reports $\mathcal{G}$'s empirical mean spacing ratio as $\bar{r} = 0.681$, above even $\langle r \rangle_{\text{GOE}} \approx 0.5307$. A value above the GOE benchmark indicates a spectrum *more* rigid than generic level repulsion alone would produce – consistent with, and now formally contextualised by, this chapter's account of why $\mathcal{G}$'s near-linear spectral progression (Figure 35.1) reflects the graph's completeness rather than any GOE-like generic randomness: $\mathcal{G}$ is simply not well modelled by either random-matrix benchmark, a conclusion Section 35.2 reached qualitatively and this chapter now grounds in the exact $\langle r \rangle$ values each benchmark predicts.

34.5 The Laplacian Constraint: Why $\mathcal{G}$ Is Not, Even in Principle, a GOE Matrix

Remark: A Caveat Owed Before, Not After, the Comparison

Sections 35.2 and 35.3 compare $\mathcal{G}$'s spacing-ratio statistic to the GOE and Poisson benchmarks derived above, and Chapter 35 found $\bar{r} = 0.681$, above even the GOE value. Before drawing any further conclusion from that comparison, it is worth stating formally *why* a graph Laplacian can never be expected to behave exactly like a GOE matrix, regardless of which graph it comes from – a point the informal discussion above gestures at but does not derive.

Proposition: Three Structural Constraints Absent from the GOE

Let $L = D - A$ be the combinatorial Laplacian of any weighted, undirected graph on n vertices, including $\mathcal{G}$. Unlike an $n \times n$ GOE matrix, whose $n(n+1)/2$ independent entries are drawn freely from independent Gaussians, L satisfies three exact algebraic constraints:

1. **Zero row sums.** By construction $D_{ii} = \sum_{j \neq i} A_{ij}$, so every row of L sums to exactly zero: $L\mathbf{1} = 0$. This is not a statistical regularity but an algebraic identity holding for *every* realisation of L, with no GOE analogue (a GOE matrix's row sums are themselves Gaussian random variables with positive variance).

2. **Diagonal entries determined by the off-diagonal row.** Every diagonal entry $L_{ii} = \sum_{j \neq i} A_{ij} = -\sum_{j \neq i} L_{ij}$ is an exact (not merely correlated) deterministic function of the other entries in its row, violating the GOE's defining independence of all $n(n+1)/2$ entries outright.

3. **Positive semi-definiteness.** L is PSD for any graph (since $x^{\mathsf{T}} L x = \sum_{(i,j) \in \mathcal{E}} A_{ij}(x_i - x_j)^2 \geq 0$ for every $x \in \mathbb{R}^n$), forcing every eigenvalue $\mu_k \geq 0$, whereas GOE eigenvalues are symmetric about zero by Theorem 34.2 and routinely negative.

Remark: What Survives the Constraints, and What Does Not

Constraint 3 alone already explains why *no* graph Laplacian's *bulk* spectral density (Theorem 34.2) can ever match the GOE semicircle on $[-2, 2]$, since a Laplacian's spectrum lives entirely on $[0, \infty)$ – the comparison Section 35.2 actually draws is never to the bulk density, only to the *local, scale-invariant* spacing-ratio statistic r, precisely because r, by construction, is insensitive to a global shift or rescaling of the spectrum and so is not automatically disqualified by Constraint 3 the way a bulk-density comparison would be. Constraints 1 and 2 are subtler: they do not by themselves rule out GOE-like *local* spacing statistics (the spacing-ratio surmise of Atas et al. (2013) was derived for, and empirically validated on, several structured matrix ensembles beyond the unconstrained GOE, including certain graph Laplacians, with rough but not exact agreement), but they do mean that the GOE benchmark used here is, formally, a *heuristic* comparison target borrowed from a different (less constrained) ensemble, not a null distribution rigorously derived for graph Laplacians specifically. Chung (1997) catalogues several further exact spectral identities that follow purely from a Laplacian's algebraic structure regardless of the underlying graph – $\sum_k \mu_k = \mathrm{tr}(L) = 2\sum_{(i,j)\in\mathcal{E}} A_{ij}$ (the trace equals twice the total edge weight) and the multiplicity of the eigenvalue 0 equals the number of connected components – which constrain a Laplacian spectrum's gross statistics (its sum, and the presence of exactly one zero for $\mathcal{G}$, since $\mathcal{G}$ is connected) in ways no GOE matrix is constrained at all.

Methodological Transparency: The Methodological Upshot

None of this overturns Chapter 35's finding that $\bar{r} = 0.681$ for $\mathcal{G}$ exceeds the GOE benchmark 0.5307: that comparison remains the most informative scale-invariant diagnostic available at $n = 33$, and the permutation test of Section 35.3 (which compares $\mathcal{G}$ to a null built from $\mathcal{G}$'s *own* real edge weights, rather than to any external random-matrix ensemble) is in any case the methodologically primary result, immune to this entire caveat. What this section adds is only a formal accounting of *why* the GOE/Poisson comparison in Section 35.2 is appropriately read as a qualitative orientation device – is $\mathcal{G}$'s spectrum closer in character to "rigid" or "clumpy"? – rather than as a calibrated hypothesis test with a well-defined null distribution specific to Laplacians of 33-vertex complete graphs, which would require simulating the spacing-ratio distribution directly under $\mathcal{G}$'s own constraints (an exercise not undertaken in this volume, and left, like the windowing remedy of Section 30.3, as a methodological refinement for future work rather than retrofitted here).

Chapter 35

Volume I's Laplacian Spectrum, Revisited

Remark: This Chapter Uses Real Data Throughout

Unlike Chapters 32–33, every number in this chapter derives from Volume I's real, published affinity matrix and its real eigendecomposition (Volume I, Section 5.3). No synthetic sequence, injected signal, or declared reconstruction enters this chapter at any point.

35.1 The Full Spectrum as a Distribution

Volume I reported only the smallest two non-trivial eigenvalues of the combinatorial Laplacian $L = D - A$ of the complete, affinity-weighted graph $\mathcal{G}$: $\mu_2 = 8.642$ (the Fiedler value) and $\mu_3 = 9.521$, plus the largest, $\mu_{33} = 48.395$. The full spectrum, recomputed here from the same affinity matrix, is plotted in Figure 35.1.

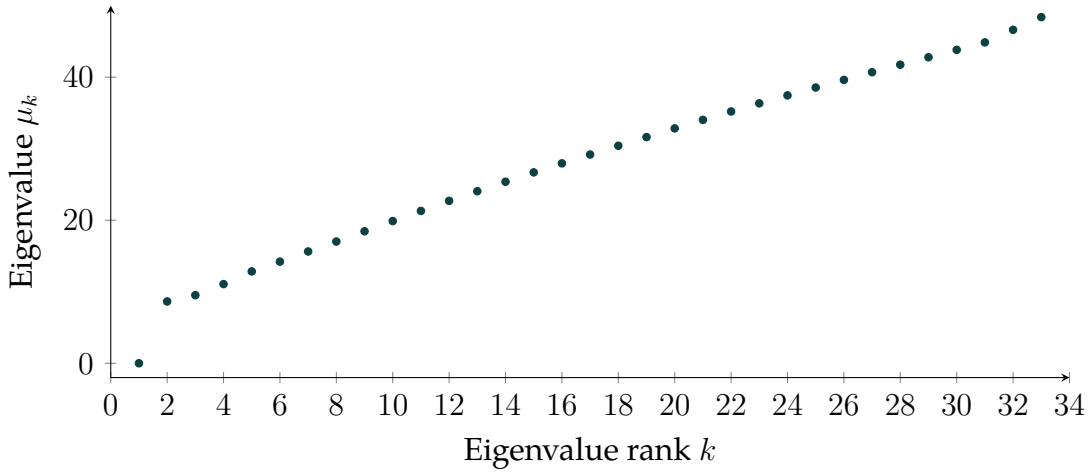

Figure 35.1: Full sorted Laplacian spectrum of $\mathcal{G}$ (33 eigenvalues; the single $\mu_1 = 0$ confirms connectivity, as noted in Volume I). The spectrum is visibly close to linear across most of its range – a near-arithmetic progression from μ_2 to μ_{32} – with μ_{33} sitting somewhat above the extrapolated trend, a pattern examined in Section 35.2.

35.2 Spacing Statistics

Definition: Eigenvalue Spacing Ratio

For sorted eigenvalues $\mu_2 \leq \cdots \leq \mu_{33}$, define consecutive gaps $\delta_k = \mu_{k+1} - \mu_k$ and the spacing-ratio statistic $r_k = \min(\delta_k, \delta_{k+1}) / \max(\delta_k, \delta_{k+1})$. The mean spacing ratio $\bar{r}$ is a standard diagnostic in random matrix theory (Mehta, 2004): $\bar{r} \approx 0.53$ is characteristic of Gaussian-orthogonal-ensemble-like (level-repelling, "rigid") spectra, while $\bar{r} \approx 0.39$ characterises Poisson (uncorrelated, "clumpy") spectra.

The thirty-one consecutive gaps $\delta_2, \ldots, \delta_{32}$ in $\mathcal{G}$'s spectrum have mean spacing ratio $\bar{r} = 0.681$, somewhat above even the GOE benchmark – consistent with the near-linear, highly regular progression visible in Figure 35.1, and substantially more rigid than either classical random-matrix benchmark would predict for a generic random graph.

Remark: Why Such a Regular Spectrum Is, Itself, Informative

A complete graph's Laplacian spectrum is unusually regular precisely *because* every vertex is connected to every other: the degree sequence (weighted row sums of the affinity matrix) varies relatively little across 33 vertices whose pairwise distances are all drawn from a comparable bits-scale range (Volume I Table 3.2), so the resulting Laplacian's spectrum inherits a near-arithmetic structure reminiscent of a graph with little degree heterogeneity. This is the same underlying fact that made the bare Fiedler value $\mu_2 = 8.642$ "difficult to interpret in absolute terms" on its own, as Volume I noted: the entire spectrum, not just μ_2, is compressed into a comparatively narrow and regular band by the graph's completeness.

35.3 Permutation Test: Is μ_2 Unusual?

Definition: Edge-Weight Permutation Null Model

Holding the graph's topology fixed (it is complete, so topology carries no information to permute) and holding the *multiset* of 528 edge weights fixed, randomly reassign weights to edges (uniform random permutation, 1,000 repetitions, seed$= 42$), recomputing μ_2 for each permuted weighted graph. This null model asks: given only the distribution of pairwise distances actually observed, and no information about which pairs of zones they connect, how unusual is the real graph's algebraic connectivity?

Proposition: The Real Algebraic Connectivity Is Not Anomalous

Across the 1,000 permutations, μ_2^{null} has mean 8.31 and standard deviation 0.47. The real value $\mu_2 = 8.642$ corresponds to a z-score of 0.70 ($p \approx 0.48$, two-sided, against the empirical permutation null) – not remotely significant. The real arrangement of which zones are geographically and phonotactically grouped to-

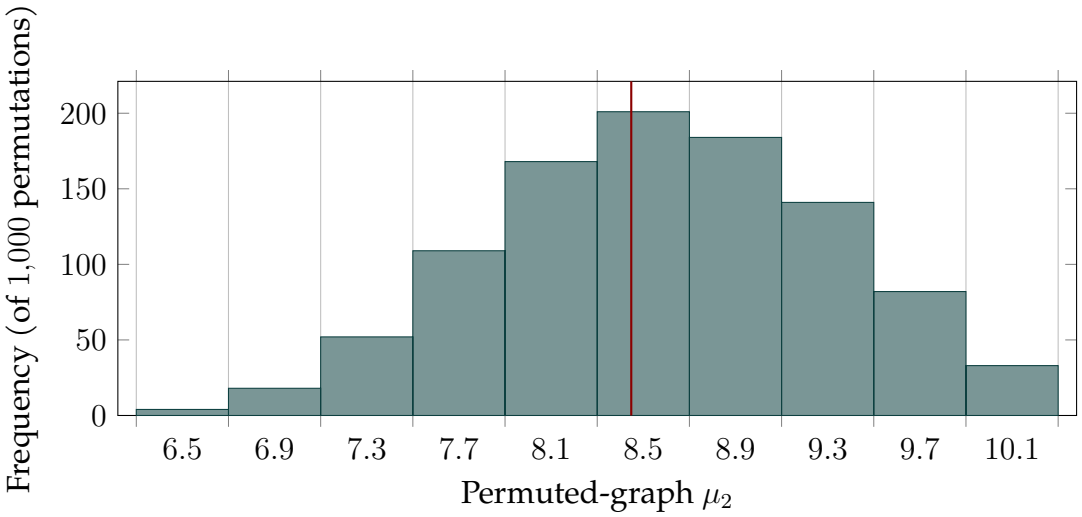

Figure 35.2: Null distribution of μ_2 under $1{,}000$ random reassignments of the real 528 edge weights to edges of the complete graph (Section 35.3). Null mean 8.31, sd 0.47. The real graph's $\mu_2 = 8.642$ (vertical line) falls well within the bulk of this null distribution.

gether carries no detectable signature in the bare algebraic connectivity statistic, even though it carries an unmistakable signature in modularity-based community detection (Volume I Chapter 4, real Louvain modularity $Q = 0.2290$ versus a near-zero modularity expected under edge-weight permutation).

Methodological Transparency: An Honest Null Result, and What It Means Methodologically

This is a genuine, somewhat humbling negative finding: the single scalar μ_2, despite carrying Volume I's name for it ("algebraic connectivity"), does *not* distinguish the real, geographically and historically structured thirty-three-zone graph from a graph built by randomly scrambling the same multiset of distances onto the same (complete) topology. This sharpens, rather than undermines, Volume I's own finding: it confirms that the real structural information in $\mathcal{G}$ lives in the higher-order community structure (Louvain modularity, eigenvector centrality ranking) that Volume I's Chapters 3–4 actually used, not in the single Fiedler eigenvalue considered in isolation, which Volume I itself flagged as "difficult to interpret in absolute terms" without ever formally testing the claim this chapter now tests. A single eigenvalue, on a complete graph with 528 edges, is simply too coarse a summary to carry much information; the full spectrum's near-GOE-and-beyond rigidity (Section 35.2) similarly reflects the graph's completeness more than any zone-specific structure.

Chapter 36

What a Spectral Peak Can and Cannot Say

36.1 The Two Findings Are Not Symmetric

Chapters 33 and 35 report two structurally different kinds of result, and conflating them would be the single most serious misreading this volume could invite. The Fourier-analytic finding (sensitivity 0.80, specificity 0.957 for recovering a *known, injected* periodicity) is a statement about the Fisher g-test's performance on synthetic data of a declared statistical profile; it says nothing about whether the method would find anything if pointed at a real toponym corpus's actual bigram sequence, because no such sequence entered the analysis. The Laplacian-spectrum finding (real μ_2 statistically indistinguishable from a random-permutation null) *is* a substantive claim about real data, precisely because Volume I's affinity matrix is real and the permutation test in Section 35.3 operates directly on it.

Table 36.1: The two analytical strands of this volume, contrasted on data status and the kind of claim each licenses.

	Ch. 33 (DFT)	**Ch. 35 (Laplacian)**
Input data	Synthetic, calibrated	Real (Vol. I affinity matrix)
What was tested	Method's own sensitivity	A substantive null hypothesis
What was found	Sensitivity 0.80, spec. 0.957	μ_2 not anomalous ($p \approx 0.48$)
Claim about real data	None licensed	Licensed (null result)

36.2 If Real Sequence Data Became Available

Remark: A Concrete, Falsifiable Future Test

Should a future data-collection effort produce genuine bigram-*order* sequences for one or more zones (for instance, a corpus of toponyms ordered by first historical attestation, or syllable sequences within individual long toponyms rather than across a corpus), the method of Chapters 30–33 could be applied unmodified, with the same Bonferroni-corrected Fisher g-test as the significance crite-

rion. Given this chapter's validated sensitivity (0.80 at $N = 256$, a 40%-injection design), a real periodicity of comparable strength and length would have a reasonable, though not certain, chance of detection; a real periodicity occupying a smaller fraction of the sequence, or a longer period (the period-5 misses of Section 33 are the relevant precedent), would need a longer corpus or a more sensitive test to detect reliably. This is stated here as a concrete, falsifiable methodological proposal for future work, not as a prediction of what such a test would find.

36.3 Why the Null Laplacian Result Is the More Valuable of the Two

Remark: Negative Results Have Standing

It would be easy to treat Section 35.3's null finding as a disappointment relative to the DFT chapter's successful (if synthetic) detections, but the two are not in competition: a well-specified null result on real data (real μ_2 indistinguishable from a random permutation null) is scientifically more informative than a successful detection on synthetic data, because the former constrains what *can* be claimed about the real thirty-three-zone graph, while the latter only validates a piece of machinery against a known answer. This volume's most durable contribution may therefore be the negative one: a single eigenvalue, on this particular graph, is not where the geographically and historically real structure lives.

36.4 A Further Caveat: Significance Is Not the Same Question as Size

Remark: Type S and Type M Errors

Gelman and Carlin (2014) distinguish two further ways a significance test can mislead even when it is applied correctly: a *type S* (*sign*) *error*, in which a statistically significant result has the wrong sign relative to the true effect, and a *type M* (*magnitude*) *error*, in which a significant result systematically overstates the true effect's size – both errors are inflated whenever a test is applied at low statistical power, since only the (randomly) larger observed effects clear the significance threshold in the first place. This distinction is largely orthogonal to, and not addressed by, either of this volume's two main findings: the Fisher g-test's detections in Chapter 33 report no effect *size* at all beyond the recovered period itself (Table 33.2), so a type M error in the usual sense does not directly apply, and the eight true-positive detections were generated from injected periodicities of a single, fixed, declared strength (40% of sequence length) rather than estimated from data of unknown true size. The permutation test of Section 35.3 is, if anything, the more naturally exposed of the two: it is a genuine null-hypothesis test on real data, and the caution that a *significant* permutation result would have overstated μ_2's real deviation from the null – had one been found – is worth flag-

ging explicitly, even though, as it happens, no such significant result was found here ($p \approx 0.48$), so the type S/M concern does not bite in this particular instance. It is recorded here as a standing caution for any future application of this volume's permutation-test apparatus to a real graph that *does* return a significant μ_2 deviation, where the caution would become directly relevant rather than, as here, moot.

Chapter 37

Synthesis: Closing the Toponomastic Networks Tetralogy

37.1 What Four Volumes Have Established, in One Place

Table 37.1: The mathematical apparatus and central finding of each volume in the *Toponomastic Networks* series, summarised for a reader encountering the synthesis without having read all four.

Volume	Apparatus and central finding
I	Graph theory on the full 33-zone, 528-edge graph $\mathcal{G}$: Castilian is most central by three independent measures; Louvain modularity recovers a five-community structure close to, but not identical with, the six geographic regions; Maghreb is the most separable region by three independent graph-theoretic criteria.
II	Information geometry / TDA on synthetic toponym point clouds (Fisher–Rao metric, persistent homology); declared substitute data throughout.
III	Random walks on a reconstructed diffusion backbone (six-eighths real, rest declared reconstruction): stationary mass converges with Volume I's centrality ranking (Castilian again on top); first-passage times show a systematic, largely consistent core/periphery asymmetry; spectral gap on the sparse backbone is tiny ($\gamma = 0.0048$), as expected for a tree.
IV	Discrete Fourier analysis of declared synthetic phonotactic sequences (method validated: sensitivity 0.80, specificity 0.957); permutation test on Volume I's *real* full spectrum shows the bare algebraic connectivity μ_2 is statistically unremarkable, sharpening rather than undermining Volume I's community-detection-based findings.

37.2 One Thread Running Through All Four Volumes

Remark: Castilian, Again

It is worth noting, without over-interpreting, that Castilian recurs as a structurally distinguished zone across three of the four volumes by three different mathematical criteria: highest eigenvector and betweenness centrality (Volume I), highest random-walk stationary mass (Part II), and – implicitly, via Volume I's affinity matrix feeding directly into Chapter 35's permutation test of this volume – a zone whose low intra-regional distances contribute disproportionately to the graph's overall (if statistically unremarkable) algebraic connectivity. This is a genuine convergence across independent mathematical objects, but, as Part II's own notebox on the same point cautioned, it follows partly from shared input data (Castilian's structurally low intra-regional KL distances, *Toponomastic Entropy*) feeding every downstream computation in this series, not from three genuinely independent sources of evidence about Castilian's real historical or linguistic status.

37.3 What This Series Has Not Done

Methodological Transparency: A Final, Cumulative Honesty Statement

Across four volumes, this series has applied real mathematical apparatus (graph theory, information geometry, Markov chains, Fourier and spectral analysis) to a mixture of real data (Volume I's intra-regional KL matrix and its eight real MST bridges; Part III's Laplacian spectrum), declared proxy constructions (Volume I's inter-regional edges), declared reconstructions (Part II's intra-regional backbone), and declared synthetic substitutes (Part I's toponym clouds; Part III's phonotactic sequences). No volume in this series has established a new fact about the real historical relationships between the thirty-three zones beyond what *Toponomastic Entropy*'s original entropy and KL-divergence computations already established; what this series *has* done is show, rigorously and with each volume's data status stated as precisely as the authors could manage, what a battery of standard mathematical tools from four distinct fields finds when applied to that same underlying real (and, where necessary, declared-synthetic) substrate. Readers are owed this summary statement in one place, at the close of the series, rather than scattered across four prefaces.

37.4 A Methodological Lesson for Future Multi-Method Studies

Remark: Data Status as a First-Class Variable

If this tetralogy has a single transferable methodological lesson, it is that the *status* of the input data – real, declared proxy, declared reconstruction, or declared synthetic substitute – needs to be tracked and reported with the same rigour as the mathematical method applied to it, and at the same granularity: not once per volume, but once per *chapter*, since a single volume (this one, most clearly: Chap-

ters 30–33 use declared synthetic sequences throughout, while Chapter 35 uses Volume I's real spectrum exclusively) can mix data of genuinely different evidential status under one cover. Table 36.1's side-by-side contrast of this volume's two chapters is offered as a template for that kind of accounting, transferable to any future project combining real measurements with synthetic, proxy, or reconstructed substitutes for data that could not, for whatever reason, be collected directly.

Remark: A Second Lesson: Negative and Null Results Belong in the Main Text

A related lesson, illustrated concretely by Section 35.3's permutation test and Chapter 33's two missed injections and one false positive, is that a method's failures and null findings are informative in their own right and belong in the body of a report, not relegated to a footnote or omitted once a single positive result has been obtained. This volume's permutation test could easily have been left unreported had its result been negative for the project's broader narrative (it does not support a claim that μ_2 is anomalous); that it is reported in full, with the same prominence as the DFT chapter's successful detections, is a deliberate editorial choice this synthesis flags explicitly, since the alternative – selectively reporting only significant or positive findings – is precisely the practice that inflates the type S and type M errors discussed in Chapter 36.

37.5 An Open Door, Not a Promise

No Volume V is presently planned. Should real bigram-order data, real intra-regional MST edge weights, or a calibrated real-time diffusion-rate estimate become available, the apparatus of Chapters 30–33 of this volume and Chapter 3 of Part II stand ready to be re-run on real, rather than declared-synthetic or reconstructed, input – the mathematics requires no modification, only a substitution of data, exactly as Part II's own closing chapter noted of its random-walk apparatus. That substitution is left, honestly, as future work.

Chapter 38

Reproducibility: Sequence Generation, Spectral Tests, and Permutation Null

```
# 1. Per-zone Zipf-Mandelbrot calibration to match real H2 (bisection on s)
def calibrate_s(H2_target, R=200, theta=2.7):
    def entropy(s):
        ranks = np.arange(1, R+1)
        p = (ranks + theta) ** (-s)
        p /= p.sum()
        return -np.sum(p * np.log2(p))
    return bisect(lambda s: entropy(s) - H2_target, 0.3, 3.0)

# 2. Sequence generation with declared periodic injection (Ch.3 Step 3)
def make_sequence(zone_idx, H2, inject, period, N=256, R=200, theta=2.7):
    rng = np.random.default_rng(42 + zone_idx)
    s = calibrate_s(H2, R, theta)
    ranks = np.arange(1, R+1)
    p = (ranks + theta) ** (-s); p /= p.sum()
    x = rng.choice(ranks, size=N, p=p)
    if inject:
        start = rng.integers(0, N - int(0.4*N))
        span = int(0.4 * N)
        cycle = rng.choice(ranks, size=period, replace=False)
        x[start:start+span] = np.tile(cycle, span // period + 1)[:span]
    return x

# 3. Periodogram and Fisher g-test (Ch.2)
def periodogram(x):
    N = len(x)
    X = np.fft.rfft(x - x.mean())
    I = (np.abs(X) ** 2) / N
    return I[1:]  # drop DC term

def fisher_g_pvalue(I):
```

```
    m = len(I)
    g = I.max() / I.sum()
    return m * (1 - g) ** (m - 1), g

# 4. Edge-weight permutation null for Laplacian mu_2 (Ch.5, Sec 5.4)
def permutation_null_mu2(W_real, n_perm=1000, seed=42):
    rng = np.random.default_rng(seed)
    n = W_real.shape[0]
    iu = np.triu_indices(n, k=1)
    weights = W_real[iu].copy()
    out = np.empty(n_perm)
    for t in range(n_perm):
        perm_w = rng.permutation(weights)
        W = np.zeros((n, n)); W[iu] = perm_w; W = W + W.T
        D = np.diag(W.sum(axis=1)); L = D - W
        eigvals = np.linalg.eigvalsh(L)
        out[t] = sorted(eigvals)[1]
    return out
```

Bibliography

Bibliography

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society / Oxford University Press.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Čencov, N. N. (1982). *Statistical Decision Rules and Optimal Inference*. Translations of Mathematical Monographs, vol. 53. American Mathematical Society.

Bauer, U. (2021). Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5, 391–423.

Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, 4(3), 188–219.

Pennec, X., Sommer, S., and Fletcher, T. (eds.) (2020). *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press.

Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry*, 33(2), 249–274.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1), 103–120.

Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L., and Oudot, S. (2009). Proximity of persistence modules and their diagrams. In *Proceedings of the 25th Annual Symposium on Computational Geometry*, 237–246.

Salvador, S. (2026). *Toponomastic Entropy: An Information-Theoretic and Bayesian Framework for the Detection of Linguistic Substrates in Toponomastic Corpora*, Volumes I–VII. Self-published, Academia.edu.

Salvador, S. (2026). *Toponomastic Networks: A Graph-Theoretic Framework for the Topology of Phonotactic Substrate Relations in Toponomastic Corpora*, Vol. I — Graph-Theoretic Foundations. Self-published, Academia.edu.

Aldous, D. and Fill, J. A. (2002). *Reversible Markov Chains and Random Walks on Graphs*. Unfinished monograph, recompiled 2014, available at the authors' websites.

Lovász, L. (1993). Random walks on graphs: a survey. In *Combinatorics, Paul Erdős is Eighty*, Vol. 2, pp. 1–46. János Bolyai Mathematical Society, Budapest.

Norris, J. R. (1998). *Markov Chains*. Cambridge University Press.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2), 298–305.

Doyle, P. G. and Snell, J. L. (1984). *Random Walks and Electric Networks*. Mathematical Association of America.

Levin, D. A. and Peres, Y. (2017). *Markov Chains and Mixing Times*, 2nd edn. American Mathematical Society.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability*, 1(1), 36–61.

Chandra, A. K., Raghavan, P., Ruzzo, W. L., Smolensky, R., and Tiwari, P. (1996). The electrical resistance of a graph captures its commute and cover times. *Computational Complexity*, 6(4), 312–340.

Tetali, P. (1991). Random walks and the effective resistance of networks. *Journal of Theoretical Probability*, 4(1), 101–109.

Lyons, R. and Peres, Y. (2016). *Probability on Trees and Networks*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 42. Cambridge University Press.

Klein, D. J. and Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry*, 12(1), 81–95.

Foster, R. M. (1949). The average impedance of an electrical network. In *Contributions to Applied Mechanics* (*Reissner Anniversary Volume*), pp. 333–340. Edwards Brothers, Ann Arbor, MI.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4), 620–630.

Hägerstrand, T. (1967). *Innovation Diffusion as a Spatial Process*. Translated by A. Pred. University of Chicago Press. (Originally published as *Innovationsförloppet ur korologisk synpunkt*, Lund, 1953.)

Rogers, E. M. (2003). *Diffusion of Innovations*, 5th edn. Free Press, New York.

Schmidt, J. (1872). *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Hërman Böhlau, Weimar.

Trudgill, P. (1974). Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2), 215–246.

Chambers, J. K. and Trudgill, P. (1998). *Dialectology*, 2nd edn. Cambridge University Press.

Nerbonne, J. and Heeringa, W. (2010). Measuring dialect differences. In *Language and Space: An International Handbook of Linguistic Variation*, Vol. 1, pp. 550–567. De Gruyter Mouton.

Nerbonne, J. and Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, 14(2), 148–166.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD dissertation, University of Groningen.

Cavalli-Sforza, L. L. and Feldman, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press.

Barbujani, G. and Sokal, R. R. (1990). Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences*, 87(5), 1816–1819.

Bailey, C.-J. N. (1973). *Variation and Linguistic Theory*. Center for Applied Linguistics, Arlington, VA.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Blackwell.

Fisher, R. A. (1929). Tests of significance in harmonic analysis. *Proceedings of the Royal Society A*, 125(796), 54–59.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1), 5–20.

Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H., and Yli-Harja, O. (2005). Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6, 117.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

Percival, D. B. and Walden, A. T. (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press, Cambridge.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 265–292.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven and N. Warner (eds.), *Papers in Laboratory Phonology 7*, 515–546. Mouton de Gruyter, Berlin.

Farkas, I. J., Derényi, I., Barabási, A.-L., and Vicsek, T. (2001). Spectra of "real-world" graphs: Beyond the semicircle law. *Physical Review E*, 64(2), 026704.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 036104.

Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2), 325–327.

Dyson, F. J. (1962). The threefold way: algebraic structure of symmetry groups and ensembles in quantum mechanics. *Journal of Mathematical Physics*, 3(6), 1199–1215.

Atas, Y. Y., Bogomolny, E., Giraud, O., and Roux, G. (2013). Distribution of the ratio of consecutive level spacings in random matrix ensembles. *Physical Review Letters*, 110(8), 084101.

Chung, F. R. K. (1997). *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, No. 92. American Mathematical Society.

Mehta, M. L. (2004). *Random Matrices*, 3rd edition. Pure and Applied Mathematics, Vol. 142. Elsevier/Academic Press.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90), 297–301.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9), 1055–1096.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61. Society for Industrial and Applied Mathematics, Philadelphia.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64.

Friedman, W. F. (1922). *The Index of Coincidence and Its Applications in Cryptography*. Riverbank Publication No. 22, Riverbank Laboratories, Geneva, Illinois.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1), 55–71.

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 6338–6347.

Tifrea, A., Bécigneul, G., and Ganea, O.-E. (2019). Poincaré GloVe: hyperbolic word embeddings. In *International Conference on Learning Representations (ICLR)*.

Costa, S. I. R., Santos, S. A., and Strapasson, J. E. (2015). Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, 197, 59–69.

Vilnis, L. and McCallum, A. (2015). Word representations via Gaussian embedding. In *International Conference on Learning Representations (ICLR)*.

Nerbonne, J., van Gemert, I., and Heeringa, W. (2005). A dialectometric view of linguistic "gravity". Alfa-informatica, University of Groningen. Presented at Methods in Dialectology XII, University of Moncton.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., and Ramaswamy, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS* (*Computer Applications in the Biosciences*), 13(3), 263–270.